Predicting the Productive Capacity of Air Force Aerospace Ground Equipment Personnel Using Aptitude and Experience Measures

THESIS
Robert S. Faneuff
Captain, USAF

AFIT/GOR/ENS/93M-05

Approved for public release; distribution unlimited

AFIT/GOR/ENS/93M-05

PREDICTING THE PRODUCTIVE CAPACITY OF AIR FORCE AEROSPACE GROUND EQUIPMENT PERSONNEL USING APTITUDE AND EXPERIENCE MEASURES

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Operations Research

Robert S. Faneuff, B.S.

Captain, USAF

March, 1993

THESIS APPROVAL

STUDENT: Capt Robert S. Faneuff CLASS: GOR-93M

THESIS TITLE: Predicting the Productive Capacity of Air Force Aerospace Ground Equipment Personnel Using Aptitude and Experience Measures

DEFENSE DATE: March 8, 1993

COMMITTEE:	NAME/DEPARTMENT	SIGNATURE
Advisor	Professor Daniel E Reynolds/ENC	
Reader	Lt Col Kenneth W. Bauer, Jr./ENS	

Preface

The purpose of this thesis was to develop experimental descriptive regression models for estimating the job performance, or productive capacity (PC), of Air Force Aerospace Ground Equipment (AGE) mechanics. The data that I used were collected under the Air Force's Productive Capacity Project by myself and other personnel from the Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory (AL/HRM), and contractor personnel from the Human Resources Research Organization (HumRRO), Alexandria, VA, and the Systems Research and Applications (SRA) Corporation, San Antonio, TX. I was fortunate enough to work with these competent people in the project planning, data collection and preliminary analyses of the collected data.

The Productive Capacity Project is part of an ongoing research and development effort aimed at identifying methods for best using Air Force personnel. The Air Force recognizes that in this day of force downsizing and shrinking defense budgets, it must make optimal use of its personnel resources. Making best use of personnel resources implies the need to be able to validly and reliably measure and quantify airmen job performance. It also implies the need to be able to model, or predict, the job performance of Air Force applicants and incumbent personnel. Mathematical modeling of job performance can contribute substantially to the Air Force's ability to better plan and use its manpower resources. This thesis research fits into the bigger picture of optimal use of resources by providing analyses of the effects of two important predictors, aptitude and experience, on airmen job performance.

This thesis would not have been possible without the continuous help and guidance from the personnel of AL/HRM. I am most indebted to Ms. Jacobina Skinner for her enormous help in gathering background material and in providing non-stop consultations throughout. I am also grateful to Mr. Bill Glasscock for his help in creating the impeccable data files that were provided

to me. And, this thesis would literally not have been possible if Lieutenant Colonel Roger Alford had not granted me permission to use the Project data. My thanks goes to all of them.

Next, I wish to express my appreciation to Professor Daniel Reynolds for serving as the thesis advisor. His continual guidance saved me from going too far astray on many occasions. And, I say thanks to Lieutenant Colonel Kenneth Bauer for his help and patience while serving as a reader and department representative for this effort.

I would be remiss if I did not thank my many classmates who helped me through this effort. In particular, I extend my appreciation to Captains Tim Mott, Randy McCanne and Tom Sterle for their unending moral support.

Last, and most of all, I wish to thank my loving and supportive wife, Mary Jean, and my children for supporting me and tolerating my absence (even when I was there).

Robert S. Fareuff

Table of Contents

List of Figures

List of Tables

Abstract

This study investigated the effects of mechanical aptitude and job experience on the job performance of 204 Air Force Aerospace Group Equipment (AGE) mechanics. Job performance was expressed as *productive capacity (PC)*, which is derived from estimated performance times on job tasks. PC measures were derived for 50 tasks typically performed by airmen in the specialty. Aptitude measures took the form of Mechanical percentile composite scores on the Armed Services Vocational Aptitude Battery (ASVAB). A second-order logistic model was used to regress PC on aptitude and experience at the task level and at the overall job, or aggregate, level. Model R^2s were generally low. For the tasks, R^2s ranged from .01 to .13, and for the aggregate model the R^2 was about .16. Generally, experience was a significant predictor but aptitude was not. There was also no indication of an aptitude/experience interaction. These results were verified through forward stepwise regression. There was some evidence that airmen may experience some skill degradation on production-type tasks at around the six year point as they transition to supervisory roles.

PREDICTING THE PRODUCTIVE CAPACITY OF AIR FORCE AEROSPACE GROUND EQUIPMENT PERSONNEL USING APTITUDE AND EXPERIENCE MEASURES

I. Introduction

1.1 General Issue.

Over the last several years, the Air Force has conducted numerous research activities aimed at developing sound ways of measuring the job performance of its personnel. These research activities were the result of three primary requirements (16:1) (30:i). First, program managers in the Air Force's manpower, personnel and training communities expressed concern that job performance measures were needed for the evaluation of their training and selection programs. Second, managers of Air Force research and development (R&D) programs needed job performance measures to serve as objective criteria for assessing the impact of various factors on individual and unit effectiveness. The third and most pressing requirement was a directive issued to the armed services in 1980 by the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics). The directive tasked the services to link their enlistment aptitude standards to job performance. This of course required the services to develop valid job performance measurement systems. Adding to the force of the directive, the House Committee on Appropriations tasked the Office of the Secretary of Defense in 1983 to provide direct oversight for joint-service research activities to address the measurement of military job performance and the linkage of job performance to enlistment standards.

These initial requirements provided the impetus for the planning and execution of several major R&D efforts by the services throughout the 1980s. These research efforts were accomplished primarily under a joint-service program called the Job Performance Measurement (JPM)/Enlistment Standards Project.

By 1990, the Air Force had developed a detailed job performance measurement system and had essentially fulfilled all the initial requirements. The Air Force, however, did not elect to abandon its research on job performance. Instead, it began the Productive Capacity Project in 1990 to continue its research on the development and potential uses of job performance measures. The Air Force felt that much more could be gained through job performance R&D. It recognized that job performance research could be of great potential value in force acquisition and manpower modeling and planning. For instance, it saw that if job performance could be modeled or predicted for those desiring to enter the service, those who would likely perform well could be identified for selection. Also, the Air Force saw that if it could model or predict the performance of its incumbent personnel, airmen could potentially be allocated or assigned to jobs so that manpower resources are best used.

The need for sound manpower modeling and planning has been highlighted by several recent events which include a virtual end to the Cold War, Operations Desert Shield and Desert Storm, sending of troops to United Nations (UN) sponsored activities, defense budget cuts, and force downsizing (30:i). There seems to be a trend of increasing world instability and a decreasing military to deal with it. What the future likely holds for the military is increasing demands placed on a smaller force. There is no doubt then that manpower resources must be planned and used wisely. This means the Air Force must be able to validly measure job performance and, more importantly, be able to predict it for its personnel.

Since the Air Force does not have a crystal ball to help it to predict the performance of if its applicants and incumbents, it has typically relied on job performance models to do the prediction. The most frequently used models have been regression-based mathematical models.

Unfortunately, the development of such models can be frustrating. Development of job performance models involves numerous elusive problems that have plagued Industrial/Organizational Psychologists and other analysts for years. For instance, developers of job performance models typ-

ically must define job performance, figure out how to accurately measure it, decide which factors influence it, and figure out how the influencing factors (called predictors) mathematically relate to performance measures—none of these have proven to be a trivial undertaking. Despite the difficulties in developing mathematical prediction models, the need for them exists, and the Air Force continues to try to develop them.

To be successful in developing mathematical models for predicting job performance, the Air Force must continue to accomplish the following items:

- Define job performance.
- Develop valid and reliable measures of job performance as defined.
- Apply the job performance measures to a representative sample of incumbent airmen.
- Identify factors likely affecting job performance (predictors).
- Look for mathematical relationships between predictor variables and the job performance measures of the airmen sample, and identify significant relationships.
- Specify an appropriate mathematical model that relates the significant predictor variables to job performance measures.
- Validate the mathematical model, perhaps on another independent sample of airmen.

It is important to point out that the above items represent a continual, iterative *process* and not a one-time-through *list*. The process can be viewed as having three distinct components or phases which are illustrated in Figure 1.

The process components are job performance measurement, job performance modeling and model validation. The process components and their subitems frequently require revisiting as more job performance measurement knowledge is gained. Each job performance research effort seems to contribute a little more to the job performance knowledge base while at the same time creating as many new research questions as it answered. Progress toward development of sound

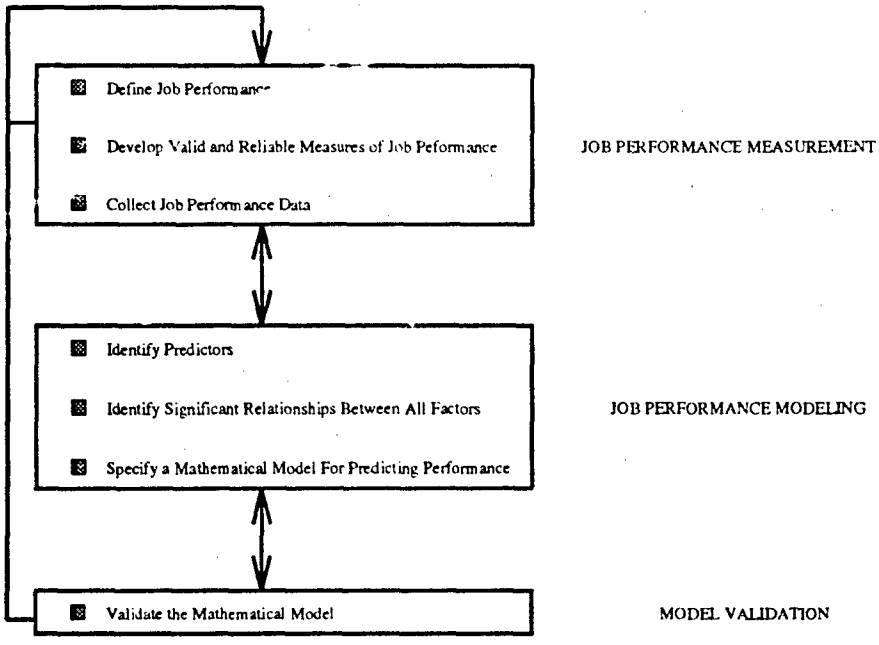


Figure 1. Job Performance Model Development Process

job performance measures and valid job performance models has been slow and has come in small increments. Much progress is yet to be made.

Whereas the joint-service JPM Project addressed mainly the job performance measurement component of the modeling process, the current Productive Capacity Project is attempting to address all of them. Initially under the Productive Capacity Project, the job performance measurement component of the modeling process was addressed—job performance was defined, experimental measures of job performance were developed, and the measures were applied to personnel in four Air Force Specialties (AFSs) (21). The next step was to proceed to the job performance modeling phase. This required identification of factors likely affecting job performance, specification of mathematical relationships between such factors and job performance, and formulation of prototype mathematical models expressing the relationship between the factors and job performance. It was this job performance modeling phase that provided the basis for this thesis.

1.2 Statement of the Problem.

The general problem facing the Air Force is that although it could greatly benefit from the ability to forecast the future job performance of applicants and incumbents, it is currently limited in its ability to do so. Development of job performance prediction models has not yet progressed to the point where current models are suitable for operational use. If suitable models are to be developed, the model development process must continue.

On a much smaller scale, the Air Force would like to use the data collected under the Productive Capacity Project to develop experimental regression-based mathematical models for relating job performance measures to certain predictor variables. The predictor variables the Air Force wishes to consider are mental aptitude and job experience.

The purpose of this thesis was to address this smaller scale problem by performing the required regression analyses on the Productive Capacity Project data to obtain the model parameter estimates needed to formulate an experimental model. In terms of the model development process expressed in Figure 1, this thesis addressed the last two items of the model development component, given the Productive Capacity Project data and the predictors, aptitude and experience.

1.3 Research Objectives.

1.3.1 Formulate a Productive Capacity Measure from Estimated Task Performance Times.

The job performance data collected under the Productive Capacity Project were in the form of estimated performance times on various job tasks specific to each of four jobs studied.

In their raw form, the estimated performance times were of limited value. One problem with them is that the raw times themselves communicate little about an individual's relative *level* of job performance. In order to assess performance level, one must first have knowledge about how others perform on the tasks, such as how long on average it takes people to do the tasks. Another problem with raw time data is that they do not have meaning outside of the associated tasks. The task

performance times have meaning only within a task within a job, and not across tasks or across jobs. Comparing performance times across tasks is like the proverbial comparison of *apples to oranges*. To illustrate these problems, consider a two-task scenario where the average times to complete the two tasks are 10 and 20 minutes, respectively. Assume an individual completes the first task in 15 minutes, and the second in 15 minutes as well. Without also considering the average performance times for the tasks, the individual's performance times suggest that performance was comparable on the tasks. But, when considering the average performance times, it can readily be seen that the individual took significantly longer than average to complete the first, and considerably shorter to complete the second. There is obviously a difference in performance levels across the two tasks that cannot be seen from the raw data.

This implies a need to standardize the performance time data. One possible standardization could be obtained through forming a ratio of the time data to a constant, say the task mean. This transformation of the time data would have the desired affect of making the resulting measure comparable across tasks. Such standardization is necessary both for making comparisons across tasks and for aggregating task-level data into overall job-wide measures that have meaning in and of themselves.

The first research objective was therefore, to find a suitable transformation of the performance time data, to standardize it. A transformation used by the Air Force in previous R&D efforts was to create a productive capacity (PC) measure from the performance time data (5) (13) (21). A PC measure is intended to express job performance in terms of how fast an airmen can perform a piece of work in reference to a standard performance time. It so happens that formulating a PC measure from the task performance times can standardize the data, giving it broader interpretability. For instance, the original PC formulation proposed by Carpenter, Monaco, O'Mara and Teachout is t^*/t, where t^* is the fastest time in which a task can be completed and t is an individual's raw performance time (5:21). With this formulation, PC always ranges from zero to one and can

be interpreted as an individual's output as proportion of maximum possible output. Other PC formulations also provide similarly helpful standardizations and interpretations.

1.3.2 Select a Task Weighting Scheme. The second research objective was to determine a weighting scheme for assigning differing levels of importance to tasks when aggregating task-level measures into overall job or aggregate measures

No weighting of the tasks implies that the performance on each task should be allowed to equally influence overall PC. This was considered a questionable practice since tasks were known to differ on such dimensions as criticality, learning difficulty, time required to perform them, and percent of time airmen spend doing them (40). Since the tasks were known to differ in importance along such dimensions, it was recognized that one or more dimensions could provide numerical values to serve as task weights that would help in better defining overall PC.

The second objective was, then, to identify an appropriate dimension from which to derive a task weighting scheme, followed by actual computation of task weights.

1.3.3 Aggregate the Task-Level Data into an Overall Productive Capacity Measure. The third objective was to determine an appropriate way of computing an individual's *overall* or *aggregate* productive capacity, using the PC measures computed at the *task* level. Task-level performance data can provide some limited insight into airmen job performance, but of ultimate importance to the Air Force is how well airmen perform *overall*. This is because Air Force jobs tend to be multifaceted requiring the performance of a variety of tasks. Jobs may also frequently change in scope. Because Air Force jobs do tend to require a variety of task skills, task-level performance data must be collapsed into overall measures that reflect an airman's ability to meet a job's overall, multifaceted demands.

The third objective was, therefore, to determine and apply a means of aggregating the task-level data into overall measures of job performance.

1.3.4 Develop Prediction Models. The fourth and most important objective was to develop descriptive regression models for relating task-level and overall PC to the predictors, aptitude and experience. The purpose of the regression models was to express how aptitude and experience appear to effect PC.

Numerous possibilities existed for the functional form of regression models. Possibilities considered included first-order and higher-order linear models, learning curve-type logarithmic models and logistic models. The objective was to select a reasonable form for the regression models, depending on the formulation of the PC measure, followed by estimation of the model parameters using appropriate techniques. As an adjunct to the research objective, the model was evaluated through residual analysis and through comparison of the model results to other performance measures and previous studies.

In short, the fourth objective was to select an appropriate regression model, estimate its parameters and analyze its results.

1.4 Scope.

Under the Productive Capacity Project, Leighton and others collected performance data on four Air Force Specialties (21). This thesis will concentrate on the analysis of data from one of these jobs, 454X1, Aerospace Ground Equipment. It was limited to the study of a single job to keep the size of the effort manageable. The methodology developed via this single-job research should find application in the analysis of the three remaining jobs by the project sponsor, the Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory (AL/HRM).

1.5 Assumptions.

Throughout this research effort it was assumed the job performance measures derived from the supervisors' task time estimates are *valid* and *reliable*. In very general terms, *valid* means

that the measures accurately measure what they purport to measure—the true job performance of the individuals studied. In equally general terms, *reliability* means that the PC measures can be *consistently* collected. Siegel and Lane (1974) describe reliability as a demonstration that measures

> do not fluctuate unduly over time as a result of something inherent in the test itself (including scorer subjectivity), the transitory nature of the function being assessed, or by factors extraneous to the particular behavior the test is designed to evaluate. (37:125)

1.6 Limitations.

A significant limitation to this thesis involves the interpretation and usability of the results. As mentioned, the goal of the thesis is to develop an *experimental* mathematical model for predicting the job performance of enlisted personnel in AFS 454X1, Aerospace Ground Equipment. The experimental model is to provide *some* insight into how the predictors, aptitude and experience, might influence an experimental measure of job performance, PC.

It must be stressed that the PC measurement methodology was still in its early stages, and the current PC measure was previously untested. Also, the model or models developed as part of this thesis include only a limited number of possible predictors. The results, therefore, are not appropriate for use in operational manpower decisions or for use in addressing any other operational concerns. The results *are* suitable for providing a basis for future research, and for providing very general ideas about how and which factors might affect job performance.

1.7 Summary.

The Air Force has recognized that it could benefit from measuring and predicting the job performance of both its current personnel and its applicants. It has undertaken several research projects with the aim of developing valid job performance measures. The Air Force's most recent R&D efforts have begun to investigate the potential uses of job performance measures in manpower and personnel decisions, and force planning and modeling.

This thesis contributes to the Air Force's R&D efforts by addressing the job performance modeling phase of the job performance model development process (see Figure 1) using data collected under the Productive Capacity Project. The remainder of this thesis documents this research. Chapter 2 provides an in-depth discussion of background material reviewed as a first step in understanding the relevant research issues. It provides an overview of the model development process, and a chronology of previous research while highlighting those items relevant to the current research objectives. Chapter 3 describes the research methodology used to prepare the data for analysis, and further describes how the regression models were estimated. It includes details of the data editing procedures, computation of aggregate PC measures, and the regression models used. Chapter 4 provides the results and pertinent discussion concerning the research findings. It provides regression results to include the estimated parameters and relevant statistics of model fit. Chapter 4 also includes correlational analyses of the model predicted values with other job performance measures. It concludes with a graphical representation of the estimated models. And finally, Chapter 5 provides a summary of the research, important conclusions and recommendations for further research.

II. Literature Review

In Chapter 1 it was explained how the primary research objective of this thesis was the development of experimental regression-based models for predicting job performance given the subjects' aptitude and experience. With this in mind, this chapter provides a background of information relevant to this thesis, couched in terms of a modeling scenario. The review thus begins with a brief overview of the modeling concept.

2.1 Overview of Modeling.

Frequently, R&D efforts involve the study and analysis of processes or systems. Such processes or systems are often very complex or not well understood. Usually the analyst desires to study a system in order to better understand it and to try to specify the relationship between system inputs and outputs.

Understanding of a system is often gained and advanced through development of a *model* representing the system. According to Law and Kelton, a model is an abstract "representation of a system developed for the purpose of studying that system." (20:3). Figure 2 depicts the relationship between an actual system and the system model. The actual system usually tends to be complex and the relationship between the inputs and outputs is usually not clearly defined or well understood. The system model attempts to clearly define the system and specify the relationships between the inputs and outputs.

It must be pointed out that not all models are good models. Some do not properly represent the system, some oversimplify the system and some can be as complex as the system itself. In general, a good model is one which is as simple as possible while still adequately representing the associated system.

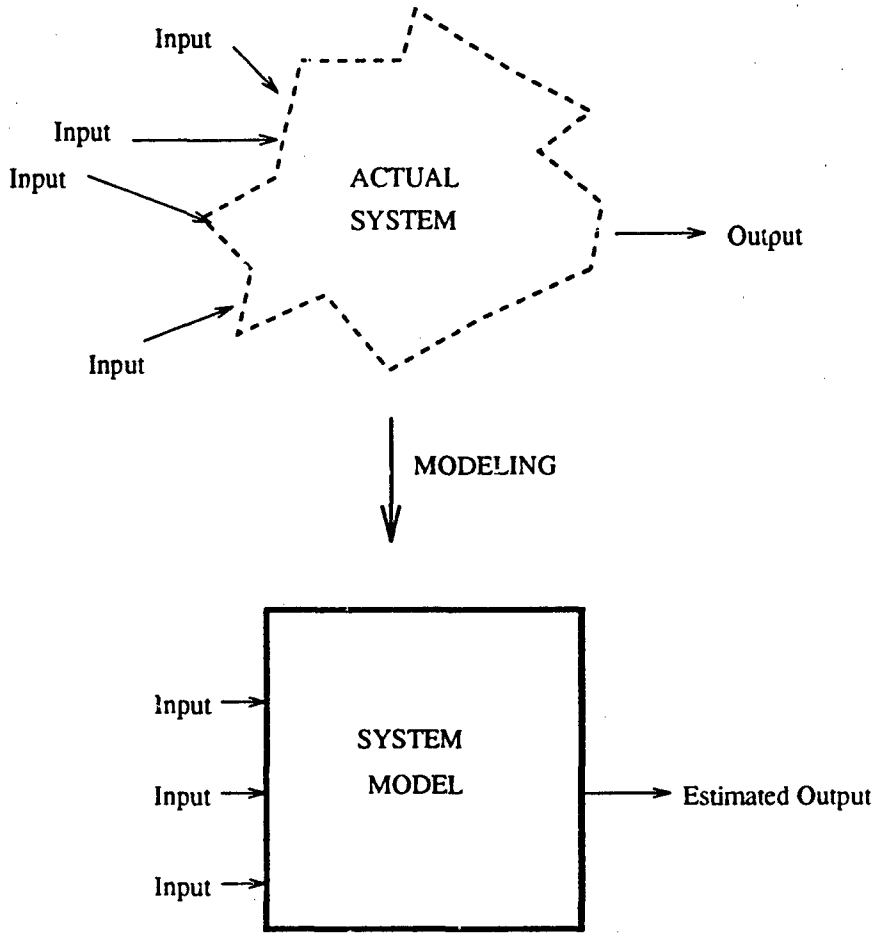


Figure 2. Graphical Representation of an Actual System Related to a System Model

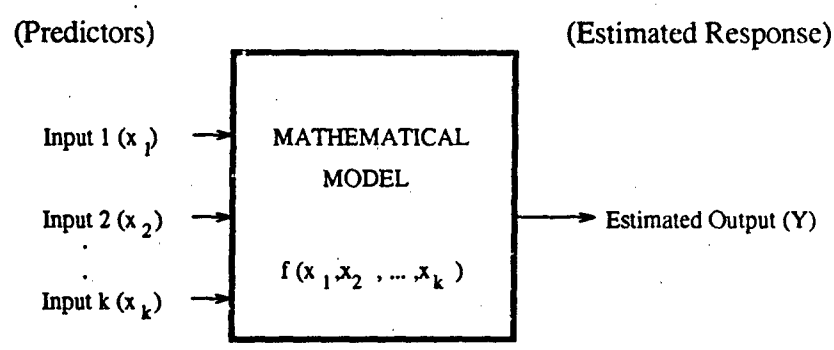


Figure 3. Graphical Representation of a Mathematical Model

There are many types of models. These include mathematical models, conceptual models, computer models, and simulation models, to name a few. Of primary concern to the current research were mathematical models because these were the type requiring development.

A mathematical model is a model in which the system is represented as a mathematical relationship between the system inputs and system outputs. In general, a mathematical model can expressed as in Figure 3.

The *black box* in Figure 3 represents the mathematical model which is generally some mathematical function of the input variables. Derivation of the mathematical function relating the inputs and output generally involves rigorous experimentation and statistical analyses to answer questions like the following:

1. Which input variables should be included?
2. Should the variables be examined in their original form, or should they be transformed?
3. How complex a model is necessary? (4:4–6)

Answering questions like those above requires application of one or more mathematical techniques. One frequently applied technique is regression analysis which is often used in analysis of linear mathematical models. Linear mathematical models and linear regression are discussed in the following section.

2.1.1 Linear Models and Linear Regression. Linear mathematical models, like all mathematical models, relate system output to inputs via mathematical functions. In linear models and most regression applications, the input variables are frequently referred to as *predictors*, and the output is often called the *response*. The predictors are often referred to as x_k, and response as Y. What distinguishes a linear model from other mathematical models is that the mathematical function relating the response to the predictors is linear with respect to the coefficients associated with the function's terms. In other words, a linear model is one that can be expressed as in Equation 1 (4:36).

$$Y = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \epsilon \tag{1}$$

where

Y	=	*the response*
$Z_1, Z_2, \ldots, Z_p$	=	*specified functions of the predictor(s), x_k*
$\beta_1, \beta_2, \ldots, \beta_p$	=	*model coefficients, or parameters*
ϵ	=	*model error terms, representing the deviation of the system data points from the underlying model.*

Note that for a model to be linear, it need only be linear in terms of the β coefficients. The functions, Z, need not be linear functions of the predictors. The Z functions are often higher-order forms of the predictors (e.g., x_k^2) or interaction terms (e.g., $x_{k-1}x_k$) to account for curvature in the curve or surface defined by the model. The actual specification of the Z functions depends on the nature of the data and the underlying mathematical relationship between the predictor and response. The analyst frequently includes various Z functions because of prior knowledge or hypotheses about the system under study. Also, preliminary analyses and data exploration can provide insight as to the specification of the Zs.

It is important at this point to make a distinction between *mechanistic* and *empirical* mathematical models (4:10-11). A mechanistic model refers to the true underlying mathematical relationship between the input and output variables. An empirical model is an approximation of the true relationship, estimated from data sampled from the system in question. Specification of the actual mechanistic model is almost always impossible or impractical due to such things as measurement and sampling error, and limited data. Therefore, it is usually the goal of the analyst to derive an empirical model using a sample of system data.

Given a linear mathematical model of the form expressed in Equation 1, *linear regression* analysis is frequently performed to aid in deriving the empirical model. Linear regression is a technique for obtaining estimates of the β parameters, given a set of predictor and response data. After estimation of the parameters, an empirical linear mathematical model can be expressed for the system in question.

Neter, Wasserman and Kutner describe regression models as serving three primary purposes (27:31). These are *description*, *control* and *prediction*. To use a regression model for description means to estimate the model parameters so that the relationship between the variables can be specified and the model can thus be used to describe the system. To use regression models for control means to specify the relationship between the predictors and response so that system specifications can be adhered to. Finally, as the name implies, prediction means the use of regression models to predict or forecast the system response given known levels of the predictors. The three purposes may overlap in a given study. It was previously mentioned that the Air Force would like to develop models for *prediction* of airman job performance. This thesis was designed to contribute to the model development process by developing regression models more for *description* than prediction. Development of such descriptive models is an integral part of the model development process as efforts are made to better understand the nature of the relationship between potential predictors and the response.

The most common method of obtaining estimates for the β parameters in linear regression is the method of *least squares*. In least squares, the model parameters are estimated such that the resulting equation they define represents a response curve or surface that minimizes the sum of the squared distances from the actual data points to the curve or surface that is estimated. Application of linear regression requires the assumptions that the values of the predictor variables for a given set of data are known constants, and also that the βs are constants that require estimation. Linear regression further assumes that the model error terms, ϵ, are independent random variables that are distributed such that they have a mean of zero. That is to say, given a fixed level of the predictors, on repeated sampling, the error is assumed to be distributed such that its mean is zero. This means that the expected value of the response, Y (denoted $E(Y)$) is $\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p$ since the βs and $Z s$ are constants.

Least squares is concerned with minimizing the squared distance between each observed Y and the its expected value, $\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p$. The equation to be minimized in least squares is expressed in Equation 2 (27:39).

$$Q = \sum_{i=1}^{n} (Y_i - \beta_1 Z_{1i} - \beta_2 Z_{2i} - \ldots - \beta_p Z_{pi})^2 \tag{2}$$

where

Q	=	*the expressed sum*
i	=	*observation number*
n	=	*total number of observations*
Y_i	=	*the response for observation i*
$Z_{1i}, Z_{2i}, \ldots, Z_{pi}$	=	*specified functions of the predictor(s) for observation i*
$\beta_1, \beta_2, \ldots, \beta_p$	=	*parameters to be estimated.*

Equation 2 is minimized with respect to the βs using standard calculus minimization techniques. The minimization yields the least squares estimators for the βs. The estimators are frequently referred to as the $\hat{\beta} s$. Least squares estimators, or $\hat{\beta} s$, have the appealing property of being minimum variance, unbiased estimators of actual βs. Having computed the $\hat{\beta} s$, the empirical regression model can be stated and the system response can be estimated, or predicted, given specified levels of the predictors. The estimated response is frequently referred to as $\hat{Y}$.

It was previously mentioned that it was assumed in linear regression, that the error terms, ϵ, were distributed for a given level of the predictors, such that their mean is zero. It is often further assumed that the ϵ not only have a mean of zero, but are normally distributed with mean zero and variance σ^2 ($\epsilon \sim N(0, \sigma^2)$). The normality assumption allows certain statistical inferences to be made concerning the regression results.

Prior to discussing statistical inferences about the regression results, the following discussion is included to show that the assumption of error terms being distributed $N(0, \sigma^2)$ implies that the $Y s$ are likewise distributed normally. This result has a direct impact on the statistical inferences which can be made. Consider Equation 1 and assume $\epsilon \sim N(0, \sigma^2)$. Since the predictors and the model parameters are constants, Y can be shown to be distributed with variance σ^2, the same variance as the error term. Since the predictors and the parameters are constants, let the right-hand side of Equation 1 be expressed as $c + \epsilon$. Next, let the variance of Y (denoted as $V(Y)$) be written as $V(c + \epsilon)$ which equals simply $V(\epsilon)$. It follows then that $V(Y) = V(\epsilon) = \sigma^2$. Further, since it is assumed that ϵ is distributed $N(0, \sigma^2)$ and that $Y = c + \epsilon$, Y not only has a variance σ^2, it is distributed $N(c, \sigma^2)$.

The assumption that the error terms, ϵ, and thus the $Y s$ are normally distributed is important when making inferences concerning the βs. Since it can be shown that the $\hat{\beta} s$ are linear combinations of the $Y s$, the $\hat{\beta} s$ are likewise normally distributed. This fact means that the t distribution can be used to make inferences about the βs. The following discussion of inferential statistics

commonly used with linear regression is an overview of the more in-depth coverage given by Neter, Wasserman and Kutner (27).

Following linear regression, it is common to test whether a given β_k is significantly different from zero. Following are the null and alternative hypotheses for such a test.

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

The test statistic t_0 is computed as $t_0 = \frac{\hat{\beta}_k}{\hat{\sigma}(\hat{\beta}_k)}$ where $\hat{\sigma}(\hat{\beta}_k)$ is the estimated standard error of β_k. The decision rule for deciding the outcome of the test is as follows.

$$\text{If } |t_0| \leq t_{(1-\alpha/2,n-p)}, \text{ conclude } H_0$$

$$\text{If } |t_0| > t_{(1-\alpha/2,n-p)}, \text{ conclude } H_a$$

Here, α represents the preselected probability of Type I error, which means α is the probability that H_a will be concluded when H_0 is true. Also, n is the number of cases on which the regression is based and p is the number of β parameters included in the model.

Enroute to discussing further statistical tests of regression results, it is necessary to introduce the concept of *sums of squares*. The sums of squares concept involves the partitioning of the sum of the squared deviations of the Ys from the average Y, $\bar{Y}$. The sum of the squared deviations of the Ys from $\bar{Y}$ is referred to as the total sum of squares and is expressed in Equation 3.

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3}$$

where

$SSTO$ = *total sum of squares*

i = *observation number*

n = *total number of observations*

Y_i = *the response for observation i*

$\bar{Y}$ = *the average response*

The total sum of squares can be viewed as a measure of total variation of the Ys from the mean response (27:87). $SSTO$ can be partitioned into two pieces, sum of squares for error and sum of squares for regression. These are expressed in Equation 4 and Equation 5, respectively.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{4}$$

where

SSE = *sum of squares for error*

i = *observation number*

n = *total number of observations*

Y_i = *the response for observation i*

$\hat{Y}_i$ = *the estimated response*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{5}$$

where

SSR = *sum of squares for regression*

i = *observation number*

n = *total number of observations*

$\hat{Y}_i$ = *the estimated response for observation* i

$\bar{Y}$ = *the average response*

The sum of squares for error represents a measure of variation of the observed data with respect to the estimated model. The sum of squares for regression represents the variation of the estimated response values with the mean response. Again, note that the total deviation of the response from the average response ($SSTO$) can be partitioned into the deviation of the observed response values from the estimated response values (SSE) and the deviation of the estimated response values from the mean (SSR). Equation 6 and Equation 7 summarize the relationship between $SSTO$, SSE and SSR (27:87–89).

$$SSTO = SSE + SSR \tag{6}$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(Y_i - \hat{Y}_i) \tag{7}$$

where

i	=	*observation number*
n	=	*total number of observations*
$Y_i - \bar{Y}$	=	*total deviation*
$\hat{Y}_i - \bar{Y}$	=	*deviation of estimated response around mean*
$Y_i - \hat{Y}_i$	=	*deviation of observed response around estimated response.*

After computation of the sums of squares, *mean squares* can be computed. The mean squares for regression (MSR) and mean squares for error (MSE) are computed by dividing the associated sums of squares by their corresponding *degrees of freedom (df)*. Degrees of freedom, in general terms, refers to the number of opportunities in which variables are free to vary, given a set of data. For instance, $SSTO$ has $n-1$ *df*, where n is the number of observations in the sample. One degree of freedom is lost because the deviations $Y_i - \bar{Y}$ must, by definition, sum to zero. This means that $n-1$ Y observations are free to vary, leaving the last observation no freedom to vary. It can be equivalently stated that one degree of freedom is lost because $\bar{Y}$ was used to estimate the true system mean (27:91). For SSE, there are $n-p$ degrees of freedom, where p is the number of parameters estimated. One degree of freedom is lost for each estimated parameter. SSR has associated with it $p-1$ *df*. There are p parameters in the model but one degree of freedom is lost because, by definition, the deviations $\hat{Y}_i - \bar{Y}$ must sum to zero. Thus, $p-1$ parameters are free to vary but the last one is not. Equation 8 and Equation 9 show the computations for MSE and MSR, respectively.

$$MSE = \frac{SSE}{n-p} \tag{8}$$

where

MSE = *mean square for error*

SSE = *sum of squares for error*

n = *the number of observations in the sample*

p = *the number of parameters included in the model*

$n-p$ = *the degrees of freedom for associated sum of squares (SSE).*

$$MSR = \frac{SSR}{p-1} \tag{9}$$

where

MSR = *mean square for regression*

SSR = *sum of squares for regression*

p = *the number of parameters included in the model*

$p-1$ = *the degrees of freedom for associated sum of squares (SSR).*

Having computed the mean squares, a common statistical test for overall regression relation can be performed. This test makes use of the fact that given the previous linear regression model assumptions, the value $\frac{MSR}{MSE}$ is distributed according to the F distribution. The null and alternative hypotheses for the test are as follows.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

$$H_a : \textit{not all the } \beta_k \ (k = 1, \ldots, p-1) = 0$$

Again, n is the number of cases included in the regression and p is the number of parameters included in the model.

Table 1. General ANOVA Table for a Linear Regression Model

Source of Variation	SS	df	MS	F_0
Regression	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F_0 = \frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$		

The test statistic F_0 is computed as $F_0 = \frac{MSR}{MSE}$ where MSR and MSE are the model mean square for error and mean square for regression, respectively. The decision rule for selecting a hypothesis is as follows.

$$\text{If } F_0 \leq F_{(1-\alpha;p-1,n-p)}, \text{ conclude } H_0$$

$$\text{If } F_0 > F_{(1-\alpha;p-1,n-p)}, \text{ conclude } H_a$$

The F test for regression relation serves primarily to determine whether any of the predictor variables, in their proposed format, are providing any statistically significant prediction of the response variable.

After computation of the sums of squares, mean squares and the F statistic for overall regression relation, linear regression results are frequently summarized with an analysis of variance (ANOVA) table. An ANOVA table is presented in Table 1.

Recall that the above statistical tests require the assumption that the error terms, ϵ, are independent random variables distributed $N(0, \sigma^2)$. The assumption of normality of the error terms is frequently tested through analysis of the *residuals*. Residual is another name for the deviation of an observed response from its predicted value, $Y_i - \hat{Y}_i$. (Residuals are often denoted as e.) Residuals are frequently analyzed to determine the aptness of the proposed regression model. The error terms ($\epsilon = Y_i - E(Y_i)$) themselves cannot be analyzed because the true mechanistic regression

model ($E(Y_i)$) is unknown and thus the error terms are unknown. To analyze the residuals, they are frequently plotted against the estimated response values and the predictor variables. These plots indicate whether the variance of the residuals (and thus the variance of the error terms) is in fact constant (σ^2) over varying levels of the other variables. Such constancy of variance is called *homoscedasticity*. A plot of the residuals against the expected residuals given a normal distribution, is also frequently plotted. This is called a normal probability plot and, as the name implies, will indicate whether the residuals (and thus the error terms) appear normally distributed.

If, after residual analysis, it appears that the estimated regression model is not apt, often either the predictor variables or the response (or both) can be mathematically transformed to make it so. Neter, Wasserman and Kutner discuss several such transformations (27).

As mentioned previously in the discussion of general modeling, not all models are good models. In linear regression, the goodness of model fit is frequently assessed through the statistic R^2. R^2 is called the *coefficient of multiple determination* and is interpreted as the proportion of variance in the response that is explained by the estimated model. The computation for R^2 is shown in Equation 10. A high R^2 indicates the estimated empirical model fits the data well and thus may provide reasonable prediction results.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{10}$$

The above tests and statistics illustrate only some of the more common descriptive and inferential statistics applied to linear regression results. While there are numerous other tests those discussed above are employed throughout this thesis and thus required review at this time.

The previous discussion of linear mathematical models and linear regression explained how model parameters are estimated and how statistical inferences can be made concerning the regression results. The previous discussion assumed that a suitable regression model was used. Con-

structing a suitable linear regression model can be a very involved process. Neter, Wasserman and Kutner describe the model-building process as involving the following four phases (27:433).

1. Data collection and preparation.
2. Reduction of the number of predictor variables.
3. Model refinement and selection.
4. Model validation.

These four phases are graphically depicted in Figure 4 (27:434). Note the relationship between Figure 4 and Figure 1 which represents the job performance model development process. Collapsing the second and third phases of Figure 4 into one phase would make the two figures highly comparable. This means that the process for developing a mathematical job performance model is virtually the same as the process for developing any linear regression model. This process can generally be extrapolated to any mathematical model development.

The *first phase* of the regression model building process, data collection and preparation, involves the gathering of the data, preferably through some designed experiment which will yield the type of data needed to answer the research questions. Following collection of the data, the data must be prepared for analysis. Data preparation may involve screening out any predictor variables which are not fundamental to the problem, which are subject to large measurement error, or which duplicate other predictors (27:435). Data preparation also involves editing of the data to remove any gross data errors, and identification of any extreme outlying observations which can adversely influence regression analyses. Useful tools for identifying data errors and outlying cases include scatterplots, histograms and frequency distributions of the predictors and response.

The *second phase* of the model building process involves the reduction of the number of predictor terms. Once the functional form of the regression relation has been decided upon (whether the predictor or response variables are to appear in linear form, quadratic form, logarithmic form,

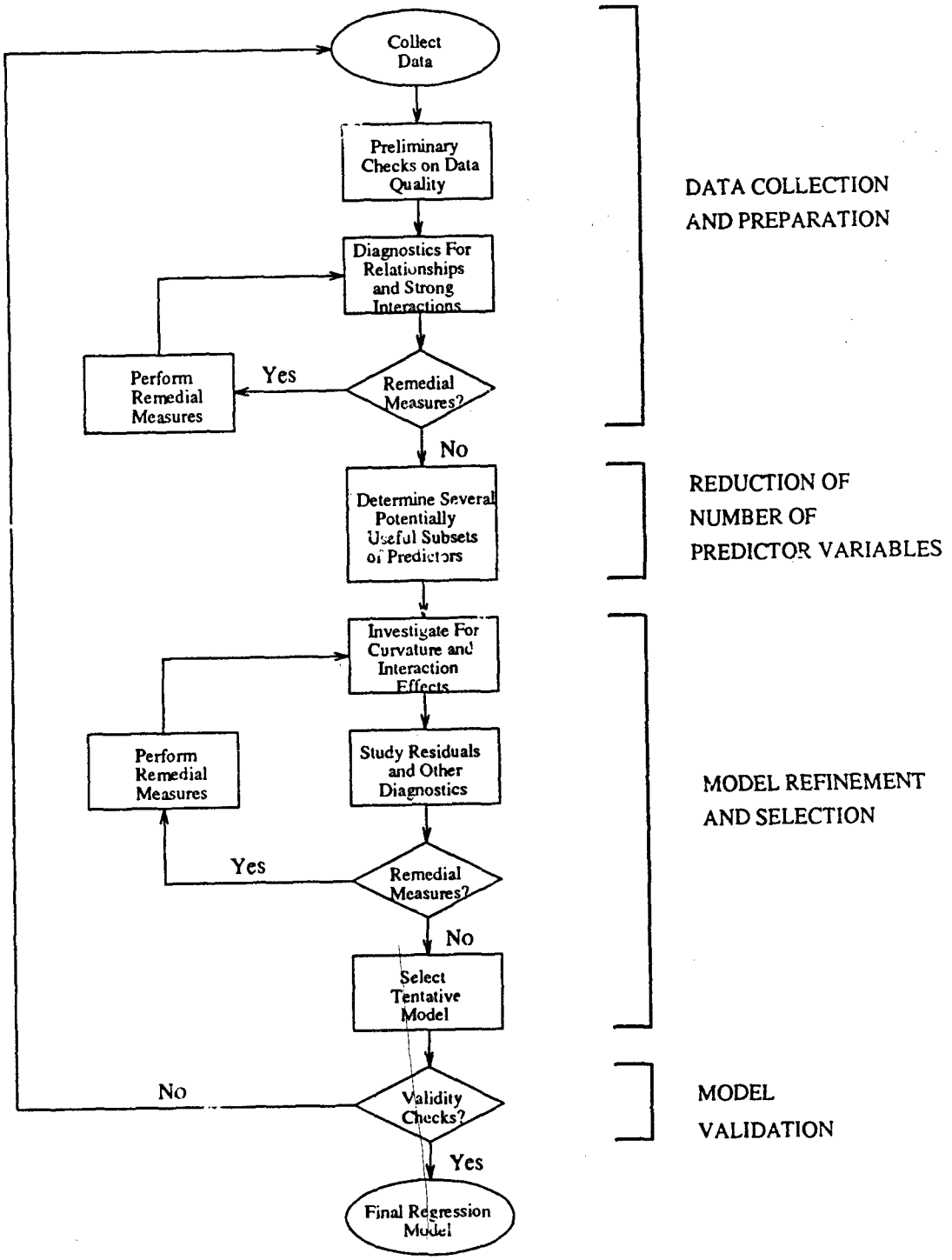


Figure 4. Strategy for Building a Regression Model

etc.), the next step is to select a good subset (or subsets) of the predictor terms (Zs) (27:43). Recall from the previous general discussion of modeling, a good model is one that not only adequately represents the underlying system, but is also as simple as possible. This is the reason for reducing the number of predictor terms, if possible. One common technique for reducing the number of predictor terms is *stepwise linear regression*. Stepwise regression is an automatic search procedure that sequentially develops the subset of predictor terms to include in the model. In very general terms, stepwise regression sequentially adds the predictor terms to the regression model and computes the statistical F test for overall regression relation. Predictor terms are added to or removed from the model based on whether their associated computed F statistics, considering other variables currently included in the model, exceed or fall below prespecified F statistic criteria (27:453–454). Stepwise regression can be an efficient way of obtaining a single, parsimonious (simple) regression model.

The *third phase*, model refinement and selection, involves study and improvement of the model(s) resulting after reducing the number of predictor terms. In this stage, the data are checked in detail for overlooked evidence of curvature and interaction effects. The model assumptions are checked through residual analysis, and diagnostics are performed to identify such things as severe outlying observations (27:437–438). Also, remedial measures such as data transformations are made if necessary. The result of this phase is the identification of a single model which most adequately and parsimoniously represents the system under study.

The *last phase* of the model building phase is model validation. Model validation involves the assessment of the model in terms of its generalizeability to the overall system, and not just to the data from which it was created. Model validation usually involves checking the model against new data, theoretical expectations, earlier results or simulation results (27:465).

Having provided a general overview of modeling with emphasis on mathematical models, and namely linear models, focus will now turn to the Air Force's most recent R&D concerning

the measurement and subsequent modeling of airmen job performance. First, however, the next section will provide the reasons that the Air Force is interested in job performance measurement and modeling.

2.2 Air Force Interest in Job Performance Research.

2.2.1 Air Force Interest in Measuring Job Performance. Aside from the need to obtain job performance measures for modeling, there are other reasons that virtually all large, success-dependent organizations are interested in measuring the job performance of their personnel. The first chapter mentioned some operational and congressionally-mandated requirements which sparked the Air Force's interest in measuring job performance. Wayne Cascio provides the following reasons that organizations in general are interested in having job performance measures(6:74).

1. Performance measures can serve as a basis for making personnel decisions such as who to fire, who to reward, and who to promote.
2. Performance measures can be used as a criteria for assessing the impact of any number of personality or situational variables on job performance.
3. Performance measures can serve as predictors of future performance.
4. Performance measures can help assess training programs and establish training objectives.
5. Performance measures can provide feedback to employees.
6. Performance measures can help in diagnosing and developing organizations.

The Air Force is interested in measuring job performance for these reasons as well. What Cascio is saying is that job performance measures can give an organization the ability to improve its manpower and personnel systems and practices in numerous ways. Coupling Cascio's reasons with operational requirements like those discussed in Chapter 1 provides the Air Force with several compelling reasons to pursue job performance measurement research.

2.2.2 Air Force Interest in Modeling Job Performance. Like *measuring* job performance, most organizations are interested in *modeling*, for prediction purposes, the job performance of their incumbent personnel and those individuals who have not yet joined the organization. The Air Force can certainly be counted among the organizations interested in modeling performance. Much can be gained by predicting the future performance of applicants. Such prediction could help in the hiring, or enlistment, process. If the Air Force could assess ahead of time who is likely to be most productive or successful, it could ensure that such individuals are enlisted, while avoiding those who are least likely to be productive.

On a grander scale, the modeling of job performance could be useful in manpower planning. Predicted job performance resulting from models, could be used as a basis for allocating personnel to various jobs according to some desired goal. For instance, if the Air Force could predict job performance, it could assign its personnel to ensure that maximum possible levels of productive capacity, or readiness, are obtained. Simply put, the ability to predict job performance can help an organization to make optimal use of its personnel resources.

2.3 Air Force Measurement of Job Performance.

Prior to reviewing relevant job performance literature, job performance models must be couched in terms of the previous modeling discussion. Figure 5 illustrates a mathematical job performance model using the same type of graphical representation shown previously. In modeling job performance, the system is in essence, a typical worker (in the current research, a typical airman). The inputs are the many factors known to influence a worker's job performance. The output, or response is the worker's actual job performance. The system (worker) is modeled as a mathematical function. In the mathematical model of job performance, the inputs are known levels of selected predictors. The output, or estimated response, is an estimate of some measure of job performance.

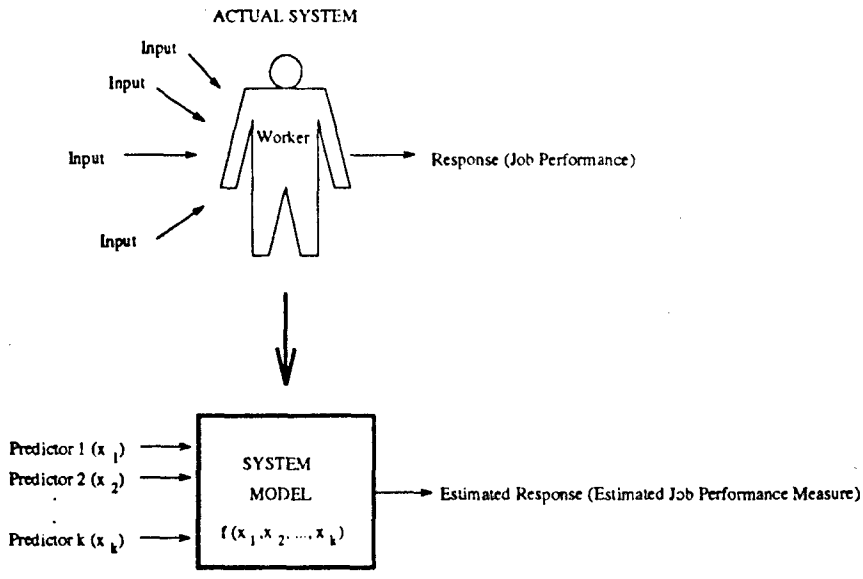


Figure 5. Graphical Representation of a Job Performance Model

It is important to note that to build a mathematical job performance model, or any mathematical model for that matter, sound measures of the response must be obtained. It should be obvious that for a job performance model, job performance is the response and job performance measures must be obtained. It will be shown that the development and collection of valid and reliable job performance measures can be a very involved process. Prior to discussing the Air Force's development of job performance measures, the next section provides an expanded definition of a job performance measure, and definitions of other key terms in job performance measurement.

2.3.1 Definitions.

2.3.1.1 Job Performance Measure. A *job performance measure* is a criterion used to assess the quality or amount of work completed.

The term *job performance measure* generally refers to the formal, valid measurement criteria used by individuals who have a professional interest in assessing and quantifying work performed on

a particular job. Job performance measures are generally not associated with informal subjective assessments or opinions concerning work completed.

Although job performance measures can theoretically be used to evaluate work done by individual workers, groups, or machines, they are usually applied to individual workers.

There are numerous possible schemes for classifying job performance measures. An important scheme that will be considered in this thesis classifies measures as *quality-based* or *quantity-based* job performance measures.

2.3.1.2 Quality-Based Measures. Quality-based job performance measures are those measures which reflect *how well* work is accomplished. Quality-based measures include such things as subjective ratings of the quality of work, or the percentage of steps performed correctly while completing a task.

2.3.1.3 Quantity-Based Measures. In contrast to quality-based measures, quantity-based measures of job performance are those that are concerned with *how much* work is accomplished, referenced to time. Some examples of quantity-based measures include the number of parts made per hour, or just the time it takes to complete a job task. It is generally the case for quantity-based measures that *more is better*. In other words, shorter work completion times are desirable. Shorter work completion times equate to higher worker output rates. This is obviously desirable for an organization provided the greater worker output is not at the expense of the worker.

The distinction between quality-based and quantity-based measures is important to the Air Force. This is because although the Air Force is interested in both quantity and quality, quantity-based measures seem to be more frequently used in most Air Force force manpower modeling and planning. The reason is that overall *work output* is usually the object of interest in planning and modeling exercises. For instance, the Air Force currently focuses on such *readiness* measures as sortie generation rates and mean time to repair aircraft. Only quantity-based job performance

measures contain the work output information needed to assess such readiness measures. But the Air Force realizes that quantity, without quality considerations, is not sufficient or desirable. Thus, there usually is a simultaneous interest in quantity and quality. This is especially true of the Air Force and most production-oriented organizations.

The Air Force's most recent job performance measurement research, under the Productive Capacity Project, has focused on quantity-based measures with an attempt to build into these measures, at least a minimum acceptable quality consideration.

2.3.1.4 Quantity/Quality Tradeoff. There is a commonly accepted notion of a *quantity/quality tradeoff* when performing work. The notion is that, all things being equal, the quality of work will decrease as the available time to put into the work decreases. Or similarly, the amount of time to complete a piece of work will increase as the attention given to quality of the work increases. It is believed that quantity and quality are directly related.

This tradeoff suggests that it is possible to allow too little time to complete a piece of work such that the quality of the work would be too low, or unacceptable. Or similarly, inordinate attention to work quality can increase the work completion time making it too long, or unacceptable. It is desirable, then, to somehow account for quantity when collecting quality-based measures, and quality when collecting quantity-based measures. This is to ensure at least minimum acceptability.

As previously mentioned, the Air Force in its Productive Capacity Project has attempted to build a minimum acceptable quality standard into its quantity-based measures. This is accomplished by phrasing the job performance measurement question as, "How long does it take to complete a piece of work *while ensuring some acceptable level of quality*?" (The actual data collection format and instrument will be discussed in Section 2.3.2.5.) For quality-based measures, quantity can likewise be accounted for by asking a measurement question like, "How well can the work be completed *in some acceptable (or fixed) amount of time*?"

The built-in quality considerations are the Air Force's current method for addressing the quantity/quality tradeoff when collecting its quantity-based measures.

2.3.1.5 Productive Capacity. *Productive capacity (PC)* is a quantity-based job performance measure that represents the maximum amount of work output a given person *is capable* of producing on a particular job or task (21).

Productive capacity is to be distinguished from *productivity*. Productivity generally refers to how much output people typically yield on a normal, day to day basis. Productive capacity on the other hand, represents the amount of work people are capable of producing if they work to their full potential.

The distinction between productive capacity and productivity is important when attempting to identify the factors affecting performance. It is quite possible, if not likely, that factors affecting productive capacity are not the same as those affecting productivity. Several recent studies have supported this theory.

The distinction between productivity and productive capacity was indirectly addressed in a study conducted by Sackett, Zedeck and Fogli (1988) (34). They made a distinction between *typical* and *maximum* performance. Typical performance generally refers to average or long term performance, while maximum performance refers to the performance resulting when maximum effort is given. Sackett and others found low correlation between typical and maximum performance of supermarket check-out clerks. Their findings suggest that a low correlation would likely exist between productive capacity, arguably a measure of maximum performance, and productivity, more a measure of typical performance. The expected low correlation implies that PC and productivity are measuring different aspects of job performance, and would likely be related to different factors.

2.3.2 Background Literature on Job Performance Measurement in the Air Force. As mentioned, several operational, practical and congressionally-mandated requirements initially gave

the Air Force motivation to rather aggressively pursue job performance measurement research. Recently, it has been the Air Force's desire to develop job performance models that has perpetuated the motivation and research. As previously discussed, measures of job performance are required in the development of mathematical job performance models. Following is review of the Air Force's research efforts to develop sound measures of job performance for meeting requirements and for development of job performance models. Two primary research projects are reviewed, the JPM and Productive Capacity Projects.

2.3.2.1 The Joint Service Job Performance Measurement Project. In response to the 1983 congressional mandate to link job performance and enlistment standards, and for numerous operational reasons, the armed services began a joint research and development project in the early 1980's. The purpose of the project was to explore valid job performance appraisal techniques. The research was coordinated across the armed services to insure a common direction of effort, to avoid duplication of effort, and to facilitate technology transfer between the services. The research project is known as the joint-service Job Performance Measurement/Enlistment Standards Project, or simply the Job Performance Measurement (JPM) Project.

2.3.2.2 The Air Force's Job Performance Measurement System Project. As part of the broader JPM Project, the Air Force began its similarly-named Job Performance Measurement System (JPMS) Project (16). As the name implies, the JPMS Project's primary purpose was to develop or identify a job performance measurement system that is valid, meaning it would consist of measures that accurately reflect how well a job is performed. As expected, this proved to be a challenging task.

The Air Force developed various job performance measures including hands-on performance tests, interviews, written tests, and supervisor, peer, and self ratings (3) (15) (16) (23). The primary performance measure developed under the JPMS Project was the Walk-Through Performance Test (WTPT) consisting of a hands-on work sample test and an interview portion (15).

The JPMS measures were eventually applied to airmen in eight Air Force Specialties (AFSs) between 1982 and 1987. The results of the JPM project are thoroughly documented by Laue, Teachout, and Harville (1992) and in numerous technical papers produced by the Technical Training Research Division, Human Resources Directorate, Armstrong Laboratory (19).

As the 1980s ended, the JPM Project, at least for the Air Force, drew to a close. However, there were no plans to operationally implement the JPM measures because of the cost and practicality problems addressed in the next section.

2.3.2.3 Problems With the Air Force's Job Performance Measurement System Measures. Despite the success of the JPMS Project in developing sound methods for measuring job performance, the JPMS measures have some problems which limit their broader use in manpower modeling. For instance, consider the Walk-Through Performance Test of the JPMS Project. Despite its attractiveness and validity as a work-sample test, it is very expensive and time consuming to develop and administer. This is because of a high degree of job and task analysis required, and because of a frequent need to access subject matter experts (SMEs), usually senior non-commissioned officers (NCOs). It also requires travel to Air Force bases for access to examinees. Further, it is intrusive in that the test must be set up and administered in the actual workplace. Finally, it requires several hours of the examinees' time, which means they must be absent from their daily duties. These factors significantly lower the utility of the measure for any kind of widespread use.

A second problem with the JPMS measures is that they are in a form that is not very useful for manpower planning (21:3). The measures are quality-based, generally in a form representing percent correct on a performance test, or a performance rating on an quality-anchored rating scale. Such quality-based measures have obscure interpretations in manpower decisions which require work output information (21:3).

Another research effort conducted during the JPM time frame introduced another job performance measure, called productive capacity, which seemed to be free of many of the troubles of the JPM measures. A discussion of this initial PC research follows.

2.3.2.4 Initial Productive Capacity Research. Carpenter, Monaco, O'Mara and Teachout (1989) conducted research for the Air Force during the same time period as the JPM Project, to explore the feasibility and utility of a novel job performance measure called productive capacity, or PC (5). As defined earlier, PC is a quantity-based job performance measure that represents the maximum amount of work output a given person is capable of producing on a particular job.

Carpenter and others mathematically defined productive capacity as t^*/t, where t^* is a standard, representing the fastest possible time in which a given piece of work can be completed. Also, t represents the time, on average, it takes the individual under assessment to complete the work.

The researchers investigated whether PC ratings could be effectively provided by Air Force supervisors. Their research involved personnel in career field 328X0, Avionics Communications.

Prior to collecting data on experimental subjects, benchmark times were assigned to clusters of tasks representative of the job, by subject matter experts. The benchmarks represented SME estimates of the average amount of time it would take a first-term airmen to complete the task cluster. The benchmarks were then provided to Air Force supervisors who used them to estimate work completion times for their personnel.

The PC data collection went as follows. Supervisors selected one of their workers whom they believed worked closest to the benchmark pace. They then estimated how long it would take each of their other workers to complete the same amount of work that the benchmark worker could perform in one hour. This was done for each task cluster. The t^* values were obtained by subtracting one minute from the fastest estimated time for each task cluster.

To validate the supervisor estimate technique, the researchers collected more objective performance data using WTPT methodology, for comparison. Correlations between supervisor ratings and the objective measures were low to moderate.

Overall, the research indicated the supervisor estimate methodology for obtaining productive capacity data had promise. This is true especially when considering the cost and time-consuming nature of empirically deriving t^* and t values by actually timing airmen while they perform job-related tasks. Unfortunately, the study indicated that more development of the productive capacity measure was needed.

This research had several associated problems. The first problem was the use of a benchmark worker as a basis for comparison when supervisors made their time estimates. Because supervisors selected unique benchmark workers (from among their own subordinates), there was to some degree, a floating reference point between supervisors when estimating performance times. This may have introduced bias into the ratings.

Second, the PC measures were computed from time estimates that were reflective of an individual's performance on average. The PC measures derived from these times do not reflect true productive capacity, but average productivity. This deviates from the definition of PC as previously expressed.

A third problem was that only a single benchmark time was used by supervisors when selecting their benchmark worker, and indirectly when making their time estimates. The single benchmark represented the average amount of time it takes a first term airman to complete work. The problem with a single benchmark is that it says nothing about the variance and distribution of performance times. This paints an incomplete picture of the range of performance times that might be expected across individuals. Supervisors probably used their own assessments of what the distribution of performance times was like and further biased the ratings.

A fourth problem was that the study looked at only one job. It is difficult to comment on the utility of the PC measure for widespread use without looking at its performance in a number of AFSs.

2.3.2.5 The Productive Capacity Project. As the JPMS Project and initial PC research drew to a close, the Air Force recognized that many operational and modeling needs for valid job performance measures would remain unsatisfied. The JPMS measures were useful in fulfilling the congressional mandate, and the initial PC research demonstrated the potential of a new measurement technique. But neither effort provided a valid and efficient measure suitable for broader use in addressing operational concerns and in development of job performance models. The Air Force realized it must conduct further research to develop a measure that could better satisfy its needs.

The Air Force reviewed its performance measurement research and determined that it would pursue the development of the PC measure over any of the JPMS measures. This is because PC offers the most overall promise. The PC measure seems to counter the problems associated with the JPMS measures in that it is relatively inexpensive to implement, it is quantity-based, and thus can be meaningful when making manpower decisions. Also, the PC measure as originally defined seemed to leave room for significant improvement.

As a result, the Air Force began its Productive Capacity Project, with the goal of improving the PC measure so that it could be used to address operational concerns and to serve as basis in manpower modeling.

The first effort of the Productive Capacity Project was an attempt to address the problems associated with the initial PC measure (21). Instead of having supervisors use a benchmark worker as a reference when estimating performance times, the researchers had them use time-anchored rating scales derived from subject matter experts as the reference. Next, supervisors were not asked to estimate individuals' typical or average performance times, but their fastest possible

performance times. And, as an alternative to providing supervisors with only a single benchmark reference time, the time-anchored rating scales used had multiple benchmarks per individual task. The benchmarks represented estimates of the *fastest* time in which the task could possibly be completed, the *average* time it would take a first term airman to complete the task, and the *longest* time that an airman would be allowed to work on the task without negative consequences to the job. Last, the researchers studied four Air Force jobs to provide a broader view of the PC measure's effectiveness.

The reader is referred to *Measurement of Productive Capacity: A Methodology for Air Force Specialties*, for a complete description of this PC research (21). Because the PC data collected by Leighton and others for AFS 454X1 will be used for the analyses in this thesis, following is a fairly detailed overview of the research.

An early issue for the Leighton and others was the selection of jobs to be studied. The first job selection consideration was the aptitude category into which jobs are classified. The Air Force uses a 10-subtest paper-and-pencil test called the Armed Services Vocational Battery (ASVAB) to select recruits for service, and then to place them into jobs. Air Force jobs can be classified into four categories corresponding to four ASVAB composite scores The job types and corresponding composite scores are Mechanical (M), Administrative (A), General (G), and Electronic (E). The composites are referred to as aptitude indices (AIs), and theoretically measure aptitude in their named area. Each Air Force job is associated with at least one AI, by the nature of the work performed in the job. There are minimum AI cutoff scores that individuals must exceed to enter the various job types (8).

To assess the utility and validity of the PC measure across a variety of jobs, the researchers opted to select one job from each aptitude area for the study. They also chose to select from among the eight jobs analyzed under the JPMS Project. This was to take advantage of the extensive task analysis information previously compiled. Also, the four jobs studied latest in the JPMS Project

Table 2. Air Force Specialties Selected for the Initial Study of the Productive Capacity Project

Specialty Code	Specialty Name	ASVAB Aptitude Index
122X0	Aircrew Life Support	General (G)
454X1	Aerospace Ground Equipment	Mechanical (M)
455X2	Avionic Communications and Navigation Systems	Electronic (E)
732X0	Personnel	Administrative (A)

Table 3. Number of Tasks Selected for the Initial Study of the Productive Capacity Project

Specialty Code	Number of Tasks
122X0	45
454X1	50
455X2	41
732X0	36

were given preference because written job knowledge tests were created for them (3). These JKTs were identified as potentially useful measures for PC validation.

A last consideration in job selection was the availability of airmen to serve as experimental subjects. After consideration of all factors, the four jobs listed in Table 2 were selected.

(Under the JPMS Project, 455X2 appeared as 328X0, Avionic Communications. The 455X2 title reflects the combination of AFSs 328X0, 328X1, and 328X4. Similarly, 454X1 appeared as 423X5.)

After selecting the jobs, the analysts were faced with the issue of selecting which tasks from within the jobs would be studied. As with the JPMS Project's Walk-Through Performance Test, the task level was chosen as the appropriate level of job detail for collecting the PC data.

Tasks from the WTPT were highly desirable candidates for the PC research because they were very well articulated and broken down into great detail as part of the WTPT development. Unfortunately, there were not enough WTPT tasks generalizable to all positions within a given AFS to provide an overall view of an individual's PC. The researchers subsequently selected additional tasks from task inventory data collected by the Occupational Measurement Squadron (OMS), Randolph AFB, TX. The final numbers of tasks are listed in Table 3.

Table 4. 454X1 Job Duty Areas

Duty Area	Description
A	Organizing and Planning
B	Directing and Implementing
C	Inspecting and Evaluating
D	Training
E	Performing General Administrative Tasks
F	Performing Preoperations or Service Inspections
G	Performing Periodic Inspections
H	Maintaining AGE Electrical or Electronic Systems
I	Maintaining AGE Engines, Motors, or Generators
J	Maintaining AGE Heating Systems
K	Maintaining AGE Refrigeration Systems or Equipment Coolers
L	Maintaining AGE Test Stand, Bomblift, or General Servicing Hydraulic Systems
M	Maintaining AGE Pneumatic Systems
N	Maintaining AGE Enclosures, Chassis, or Drives
O	Maintaining Mobile Tactical Air Control Systems Equipment
P	Dispatching AGE
Q	Maintaining Special Tools or Shop Equipment
R	Performing Quality Assurance Tasks
S	Performing Nonpowered AGE Maintenance
T	Performing Cross-Utilization Tasks

Job tasks are typically coded by OMS (40). Task codes consist of a letter prefix and a numeric suffix. The letter prefix identifies which job duty area the task is from, and the numeric suffix differentiates tasks within the duty areas. Because data for AFS 454X1 were analyzed in this thesis, Table 4 which lists the 454X1 job duty areas and Table 22 at Appendix A which lists and describes the 50 454X1 tasks analyzed were included (40).

The task descriptions in Table 22 at Appendix A do not exactly match the descriptions maintained by OMS. The task descriptions had to be modified for the Productive Capacity Project to clearly define a task by specifying exact equipment and precise starting and stopping points so that accurate completion time estimates could be made.

After task selection, the researchers had to establish benchmark times for the tasks. The benchmarks were needed for the creation of the rating scales to be used by the supervisors in

estimating the work completion times of their subordinates. Three benchmarks were derived for each task. These represented the fastest time in which the task could be completed, the average time it takes a first term airman to complete the task, and the longest time that an airman would be allowed to work on the task without significant consequences to the job.

To get these benchmarks, six SMEs from each job were assembled for workshops at Brooks AFB, TX. The workshops for each job were held separately. During the workshops, the SMEs were presented the task lists corresponding to their given jobs. The Nominal Group Technique (NGT) was used to reach consensus among the SMEs for each benchmark for each task (14).

A detailed analysis of the interrater agreement of the SMEs when providing the benchmarks was accomplished by Skinner, Faneuff, and Demetriades (1991) (39). Overall, they found that there tends to be very strong agreement among SMEs when estimating the benchmarks.

To gain access to supervisors and airmen to serve as experimental subjects, it was necessary for the researchers to visit a number of Air Force bases. The primary considerations in selecting the Air Force bases included the following:

- The number of potential subjects available at each base
- Base location (Continental U.S or overseas)
- Base mission (training, classified, etc.)

The researchers determined that 10 bases would be visited. The bases are listed in Table 5.

A sample size of 200 airmen per AFS was targeted. This was the maximum number that could be tested given project resources. Also, a sample size of 200 was considered sufficient to support planned analyses. Subjects for each AFS were selected to be representative of the base populations in terms of three factors:

- Job experience

Table 5. Bases Visited in the Initial Study of the Productive Capacity Project

Air Force Base
Travis AFB, CA
Beale AFB, CA
George AFB, CA
Davis-Monthan AFB, AZ
Holloman AFB, NM
Langley AF3, VA
Shaw AFB, SC
Offutt AFB, NE
Eglin AFL, FL
Charleston AFB, SC

- Race
- Gender

Job experience was considered a very important factor because of its hypothesized statistical relationship with PC. Because of the hypothesized relationship, an attempt was made to get subjects across a range of experience. This would allow the hypothesis to be appropriately tested.

Experience was expressed in terms of skill level. Skill level is a variable used by the Air Force. It ranges from 0 to 9, and it represents the amount of training, expertise, and experience an airman has on a given job. Skill levels 3 and 5 were sought because they indicate that an airman is performing mostly hands-on production work, as opposed to receiving technical training or performing supervisory duties. Race and gender factors were considered important to allow for future investigation of differential effects of the PC measure across race and gender groups.

The researchers reviewed distributions of personnel at the participating bases and developed target numbers of subjects. The actual individual test subjects were selected by the participating bases, using guidance from the researchers. The bases had to select the subjects because they had the most current information on manning requirements, deployments, and personnel status. One problem with having the bases select the subjects, was that no consideration could be given to subject aptitude level. This is because ASVAB scores were not available in base-level personnel

Table 6. Sample Sizes for the Initial Study of the Productive Capacity Project

Specialty Code	Number of Subjects
122X0	159
454X1	204
455X2	155
732X0	193

records. Like experience, aptitude is expected to be related to PC and it would have been desirable to sample subjects from a range of aptitude levels. The final sample sizes are listed in Table 6.

The primary focus of Leighton and others' research was to collect appropriate data to allow them to assess how well supervisors can *estimate* the task completion times of their subordinates.

This means that the primary measurement instrument of the study was the time estimation forms and accompanying booklets used by the supervisors to estimate how long it would take their subordinates to complete the tasks being studied (24). The rating forms and booklets provided the supervisors with detailed task descriptions, and a time line showing the fast, normal, and slow times for task completion.

It was on the estimation forms that the supervisors provided the task completion time estimates, as well as an indication of how frequently they have seen the ratee complete the task (Regularly, Often, Never). In making their time estimates, supervisors were told to "think about how long it would take each airman to do the task if he or she were working as quickly as they could, while maintaining satisfactory performance" (21:52).

In addition to using the forms to estimate task performance times, the supervisors used them to provide an overall or global estimate of their subordinates' productive capacity. The supervisors were asked to answer the following question: "In this specialty, consider the maximum amount of acceptable work that can be done by a person in a typical day as 100 percent. What percent of the maximum could the person you are currently rating do in a typical day?" (21:52) This measure was

of secondary interest, and was collected for use as an object of comparison for the time estimates. Figure 6 provides an example of the time estimation form for 454X1.

Besides the PC rating forms, many other instruments were used during the study. These instruments are to be used to validate the PC estimation methodology, and to investigate for relationships between scores from these instruments and PC. Other data forms were used to collect background information on the experimental subjects and their rating supervisors.

The main instrument for validating the PC estimation methodology is a hands-on test similar to the hands-on portion of the WTPT developed under the Air Force's JPMS Project. For the test, a relatively small subset of tasks was chosen from each job (between 8 and 11). A subsample of the experimental subjects were then chosen to actually perform the tasks (60 airmen from each AFS). As the subjects performed the tasks, the researchers used a stopwatch to determine their performance times. This was determined to be the best possible way to validate the supervisor estimates. Also, JKTs were administered to subjects in three of the four jobs studied (none was available for 455X2). JKTs are written tasked-based, multiple-choice tests designed to measure how well an airman knows the procedures required to perform job tasks (3). The JKTs are to serve as a basis of comparison in which to evaluate the PC measure. In previous studies, corrected correlations between the hands-on portion of the WTPT and JKTs were found to be between .50 to .80 indicating a moderate to high level of linear relationship (19:11). Since the estimated PC measure in the current study and the JKT are both purported to measure job performance, it was expected that these measures would be correlated to some degree as well. High correlation was not expected because the instruments likely measure different dimensions of performance since the JKT deals with *how well* an individual knows the job, and PC deals with *how long* it takes an individual to do work on the job.

Other measures that were administered include a 160-item interest inventory, the *VO*cational *I*nterest For *C*areer *E*nhancement (VOICE), which was administered to subjects to determine their

AFS 454X1

AEROSPACE GROUND EQUIPMENT (AGE) SPECIALIST

Name_____

Airman's Name _____ Airman's SSN_____

R=Regularly
O=Occassionally
N=Never

Hr=Hours
Min=Minutes
Sec=Seconds

How Often You Observe Incumbent Perform Task (Check One Box)	Consensus Performance Time Scale			What is Incumbent's Performance Time? (Fill in Box Below)
	← Fastest	Normal	Slowest →	
E120 R O N	8 min	13 min	21 min	Hr Min Sec
F154 R O N	11 min	17 min	24 min	Hr Min Sec

In this specialty, consider the maximum amount of acceptable work that can be done by a person in a typical day as 100 percent. What percent of maximum could the person you are currently rating do in a typical day? Write your estimate in the box below.

1% ——— 100%

Percent of Maximum

_____ %

Figure 6. Example of the 454X1 Rating Form

level of interest in their current job (9) (12). A 30-item motivation measure, the Generalized Motivation Scale (GMS), was also administered to the subjects (32) (33). The GMS was administered to allow for investigation of any relation between overall motivation and PC.

Numerous other data collection forms were used to gather background information on both subjects and supervisors.

The project sponsor, AL/HRM, has assembled a rich data base made up of individual subject records containing the project measures described above. Adding to its value, data from other important Air Force files have been added to ensure a complete background on the experimental subjects. Data added from the Uniform Airmen Record (UAR), a periodically-updated file maintained by the Air Force Military Personnel Center (AFMPC), included education level, race, ethnicity, and the date in which the subject began active service. Data from Military Entrance Processing Station (MEPS) files included aptitude scores and other background information for cross-checking purposes.

2.4 The Predictors of Job Performance.

In the previous section, significant discussion concerning the job performance model response, job performance, was provided. Next, discussion focuses on the predictor variables. Recall that the predictor variables to be used in this thesis are aptitude and experience.

Numerous factors are thought to influence the job performance of individuals. These include personality traits, job satisfaction, job interest, aptitude, and experience, to name just a few. Psychological research is filled with studies showing the effects of such factors on performance.

It is important to note that the job performance measure under study in this thesis is productive capacity, which is distinguished from productivity. Many individual attributes that influence productivity like job interest, motivation, and other personality factors were not expected to influence productive capacity because PC is a measure of a person's *capacity to produce* not their *actual*

Table 7. ASVAB Subtests

Subtest Name	No. of Items	Testing Time (Min.)
General Science (GS)	25	11
Arithmetic Reasoning (AR)	30	36
Word Knowledge (WK)	35	11
Paragraph Comprehension (PC)	15	13
Numerical Operations (NO)	50	3
Coding Speed (CS)	84	7
Auto-Shop Information (AS)	25	11
Mathematics Knowledge (MK)	25	24
Mechanical Comprehension (MC)	25	19
Electronics Information (EI)	20	9

production. PC is theoretically independent of how a person views the work or feels about the job. PC however was not believed to be independent of such things as a person's mental aptitude or job experience since these likely influence a person's capacity to produce. Because of the hypothesized relationships between aptitude, experience and productive capacity, the Air Force's emphasis has been on aptitude and experience as predictors of PC (5) (13). This thesis continued with the analysis of aptitude and experience as predictors.

In Air Force studies, aptitude is usually expressed in terms of scores on the ASVAB. As previously mentioned, the ASVAB is a 10-subtest, paper-and-pencil test given to all armed service and Coast Guard applicants (10). The test is designed to measure aptitude in various areas. The applicants' ASVAB scores determine whether or not they are selected for service, and if so, what type of job they are classified into (10).

The Air Force uses five ASVAB composite scores to select and classify applicants and recruits. Table 7 and Table 8 show the ASVAB subtests and composites, respectively, used by the Air Force.

The ASVAB is validated against a number of criteria by each of the services. The Air Force typically uses the final grades Air Force recruits receive in technical training schools as validation criteria. For instance, Ree and Earles (1992) accomplished an ASVAB validation study in which they analyzed data from 88,724 Air Force recruits completing 150 training courses (31). For 22 jobs

Table 8. ASVAB Composites Used by the Air Force

Composite Name	Definition
Armed Forces Qualification Test (AFQT)	$2VE + AR + MK$
Verbal (VE)	$WK + PC$
Mechanical (M)	$MC + GS + 2AS$
Administrative (A)	$NO + CS + VE$
General (G)	$VE + AR$
Electronic (E)	$AR + MK + EI + GS$
MAGE	$M + A + G + E$

The composites are computed using subtest standard scores.

which use the M composite for selection, the corrected-for-range-restriction correlation coefficients between the M composite and final school grades ranged between .63 and .78. For 11 jobs which use the A composite, the correlation coefficients between A and final school grades ranged from .58 to .74. For 52 jobs using G, correlation coefficients ranged from .04 to .85. And for 44 jobs using E, the correlation coefficients ranged from .56 to .90 (31:11–13). These moderate to high correlation coefficients tend to indicate the ASVAB is valid, at least for predicting training school success.

This has long been the Air Force's method of choice for validating the ASVAB, but it is recognized that validating the ASVAB against training grades does not necessarily equate to validating the ASVAB against job performance. But, studies by Carpenter and others, Faneuff and others, and AL/HRM indicate that ASVAB scores can potentially be a significant predictor of PC, a job performance measure (5) (7) (13) (38).

Experience measures in Air Force job performance R&D are usually expressed in terms of total months of active federal military service (TAFMS). This is generally used as a surrogate for job experience because job experience indicators are not readily obtainable from existing computer files. The reason job experience is considered important as a predictor can be traceable to learning curve theory. Learning curve theory basically states that the time it takes to complete a unit of work will decrease as the operator becomes more experienced (41). This suggests that PC will likewise be affected by job experience because PC is computed from performance time data. As a result, experience is an important predictor in PC prediction models.

2.5 *The Relationship Between Job Performance, Aptitude, and Experience.*

The previous sections discussed the response, job performance, and the predictors, aptitude and experience. This section discusses how job performance has been shown to relate to the predictors in previous modeling efforts.

To analyze the effects of aptitude and experience on job performance Schmidt, Hunter, and Outerbridge (1986) performed a study based on a sample of 1,474 civilian and military personnel (35). They used path analysis to analyze the impact of job experience and mental ability on job performance. The measures of job performance used were written job knowledge tests, work sample tests, and supervisory ratings of job performance. Their findings suggest that job experience affects job performance in two ways. First, greater job experience indirectly effects performance because it leads to greater acquisition of job knowledge. The greater job knowledge leads to greater performance. Second, job experience directly affects the ability of people to perform work-related activities as indicated by work sample tests. Mental ability was found to have the same pattern and magnitude of relationships on job knowledge and work sample performance as experience.

Schmidt, Hunter, Outerbridge and Goff (1988) conducted a study based on the same sample as the previously cited study, to analyze the joint relation of experience and mental ability with job performance (36). They tested three hypotheses. The first, the *divergence* hypothesis, "predicts that as job experience increases, the performance difference between high- and low-ability employees will increase." (36:46) The second, the *convergence* hypothesis, "proposes that as employees gain job experience, initial ability becomes less important as a determinant of job performance." (36:46) Last, the *noninteractive* hypothesis states "experience increases job performance of high- and low-ability employees at the same rate." (36:47) In other words, the third hypothesis states that there is no interaction between experience and ability. Their findings support the noninteractive hypothesis, and that mental ability and experience are important determinants of job performance.

In a similar study, Alley and Teachout (1990) used the WTPT data collected during the JPMS Project (1). Like Schmidt, Hunter, Outerbridge and Goff, their findings support the noninteractive hypothesis and the fact that mental ability and experience are important determinants of job performance.

2.6 Air Force Job Performance Modeling Research.

The preceding sections provided an overview of linear models and discussions of the response, PC, and the predictors, aptitude and experience. The stage has thus been set for discussion of specific Air Force studies in which the response variable was PC or raw performance times, and the predictors were aptitude and experience.

To model PC, Carpenter and others used the logistic growth model in Equation 11 to model PC (5:21).

$$PC = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2}} + \epsilon \tag{11}$$

where

PC	=	*productive capacity*
x_1	=	*experience (months in Air Force)*
x_2	=	*ASVAB aptitude score (Electronic Composite)*
$\beta_0, \beta_1, \beta_2$	=	*parameters to be estimated*
ϵ	=	*the model error terms.*

Note that the logistic model in its original form was not a linear mathematical model because it was not linear with respect to the β parameters. However, the logistic model was linearized for application of linear regression.

Carpenter and others linearized the logistic model by making the transformations indicated in Equation 12 (5:21-23). Linearizing the model equation as such allowed for estimation of the model parameters using least squares estimation.

$$ln(\frac{PC}{1-PC}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{12}$$

where

$ln(\frac{PC}{1-PC})$ = *the logit of productive capacity*

x_1 = *experience (months in AFS)*

x_2 = *ASVAB aptitude score (Electronic Composite)*

$\beta_0, \beta_1, \beta_2$ = *parameters to be estimated*

ϵ = *the model error terms.*

Using Equation 12, Carpenter and others modeled PC at the task cluster level, and also at the aggregate or overall level. The aggregate measure predicted was computed from a weighted average of task cluster performance times. They analyzed a total of 10 task clusters. Across the 10 task clusters, the estimated experience coefficient was significantly different from zero at the $\alpha = .05$ level in seven cases, and the estimated aptitude coefficient was significant in four cases (5:22). Model R^2s ranged from .00 to .39 across the clusters. And, the models showed significant regression relations in eight cases. For the aggregate model, both the estimated experience and aptitude coefficients were significant at the $\alpha = .05$ level. The aggregate model R^2 was .44 and the model regression relation was significant at the $\alpha = .05$. Overall, the results suggest the supervisor estimate method for generating individual performance times has potential. But, as Carpenter and others point out, further refinement is needed (5:51)

While Carpenter and others used the logistic model for predicting PC, Faneuff and others found that a linear model provided better model fit than did the logistic model (13:9). Faneuff and

others estimated PC at the overall or aggregate level using the model expressed in Equation 13 (13:9-10). Faneuff and others computed PC as *WTPT score/maximum observed WTPT score*, using data collected under the Air Force's JPMS Project.

$$PC = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{13}$$

where

PC	=	*productive capacity*
x_1	=	*ASVAB aptitude score*
x_2	=	*experience (months in Air Force)*
x_3	=	*a binary variable representing skill level (coded 1 if skill is 5 or higher, 0 otherwise)*
$\beta_0, \beta_1, \beta_2, \beta_3$	=	*parameters to be estimated*
ϵ	=	*the model error terms.*

The model was estimated for six of eight jobs studied under the JPMS Project. One job, Aerospace Ground Equipment (the job studied in this thesis), was analyzed using two ASVAB aptitude composites, Electronic and Mechanical, yielding a total of seven possible prediction models. The regression results showed a significant aptitude coefficient in four of seven total cases, a significant experience coefficient in four cases, and a significant skill level coefficient in three cases (all coefficients were tested at the $\alpha = .05$ level). Model R^2s ranged from .10 to .23 (13:9-10).

AL/HRM modeled estimated performance time data (as opposed to PC data) at the task level, using the learning curve model expressed in Equation 14 (7) (38). The data used was that collected by Leighton and others for the Aerospace Ground Equipment specialty (the same data used in this thesis) (21).

$$ln(t) = \beta_0 + \beta_1 ln(x_1) + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad (14)$$

where

t	=	*estimated task performance time*
x_1	=	*experience (months on the job)*
x_2	=	*ASVAB aptitude score (Mechanical composite score)*
$\beta_0, \beta_1, \beta_2, \beta_3$	=	*parameters to be estimated*
ϵ	=	*the model error terms.*

A general form of the learning curve is expressed in Equation 15 (2) (18). Like the logistic model used by Carpenter and others, the learning curve model in its original form is not a linear model. But, like the logistic model, the learning curve model can be linearized so that its parameters can be estimated via least squares. The linearized learning curve model is expressed in Equation 16. Note that AL/HRM's linearized model (Equation 14) is analogous to the general form of the linearized learning curve model (Equation 16). A typical learning curve is plotted in Figure 7.

$$t = Ax^{\beta_1} + \epsilon \quad (15)$$

where

t	=	*task performance time*
x	=	*units of experience*
A, β_1	=	*parameters to be estimated*
ϵ	=	*the model error terms.*

Equation 15 can be written in linear form as Equation 16.

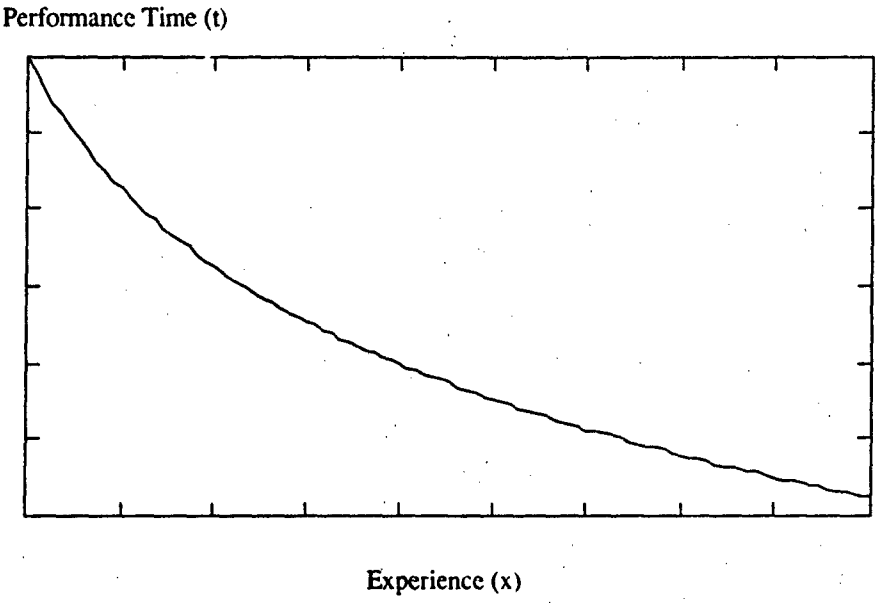


Figure 7. Plot of a Typical Learning Curve

$$ln(t) = \beta_0 + \beta_1 ln(x) + \epsilon \tag{16}$$

where

t = *task performance time*

x = *units of experience*

β_0 = $ln(A)$, *a parameter to be estimated*

β_1 = *a parameter to be estimated*

ϵ = *the model error terms.*

Using the linearized learning curve model expressed in Equation 14, AL/HRM found significant coefficients for *ln(job experience)* for 26 of the 50 tasks, significant aptitude coefficients for 18 tasks, and significant *aptitude* × *experience* interaction coefficients for 14 tasks (all coefficients were tested at the $\alpha = .05$ level) (38). Model R^2s ranged from .01 to .20. The models showed significant

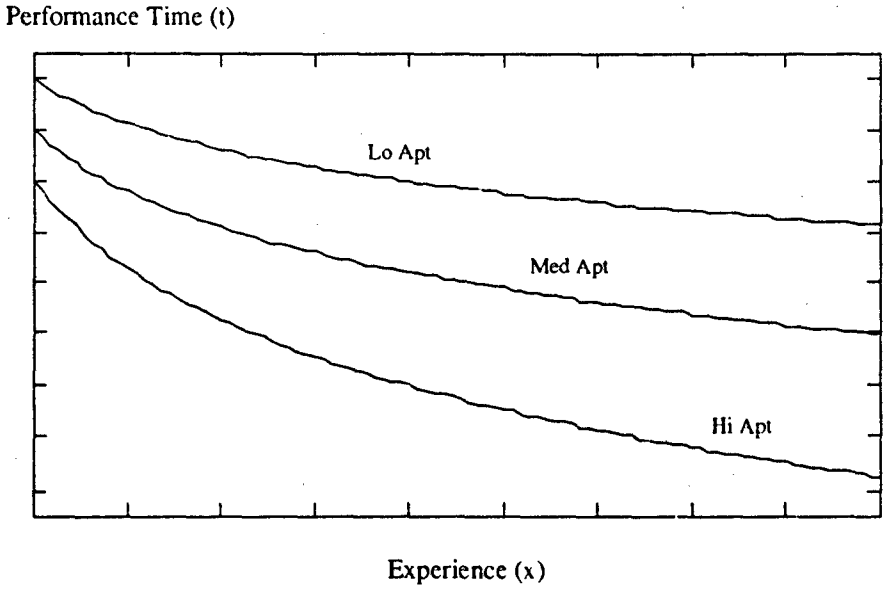


Figure 8. Plot of Learning Curves Broken Out by Aptitude

regression relations at the $\alpha = .05$ level for 41 of the 50 tasks. The inclusion of the aptitude and interaction terms in their model allowed AL/HRM to create learning curves broken out by aptitude group. An example of learning curves broken out by aptitude groups can be found in Figure 8.

One significant problem with the learning curve model is that there is no apparent way to model job performance at the overall job level. To model overall job performance using the learning curve model would likely require aggregation of task-level performance times. Such an aggregated measure would have a dubious interpretation.

2.7 Relating the Literature to the Research Objectives.

The previous sections of this chapter outlined modeling in general, and a great deal of literature on the Air Force's job performance modeling R&D. This section serves to provide a brief overview of the literature with specific reference to the research objectives outlined in Chapter 1.

2.7.1 Formulating a Productive Capacity Measure from Estimated Task Performance Times. As previously mentioned in Chapter 1, an important research objective of this thesis was to identify an appropriate way of formulating a PC measure from the estimated task performance times collected under the Productive Capacity Project (21). This PC formulation was to transform the estimated performance times into a standardized measure that has meaning across tasks and across jobs. Recall that such standardization was also required for aggregating data across tasks. There are probably numerous ways that the estimated task performance times can be transformed into a meaningful PC measure. In the literature, three ways to formulate a PC measure were proposed. These are described below.

2.7.1.1 The Formulation of Productive Capacity During the Initial Productive Capacity Research. In one of the first PC research efforts, Carpenter and others did a study of Avionics Communications Specialists and proposed the original PC formulation. They computed PC as t^*/t, where t^* is the fastest possible time in which a given amount of work can be completed, and t is the time that it takes the individual being assessed to complete the work (5:21). In this original work, the t^* and t measures were applied to clusters of tasks.

This formulation has the desirable quality of ranging from zero to one, which results in an intuitively appealing interpretation. The measure can be interpreted as an individual's output as a proportion of maximum possible output.

Although the data collected under the Productive Capacity Project is collected at the task level as opposed to the task cluster level, the measure could just as easily be formulated at the task level, as in the case of the current research.

Faneuff and others used an adaption of the PC formulation of Carpenter and others in an effort to extend Carpenter and others' work to a greater number of jobs (13). The formulation used by Faneuff and others was in fact t/t^*, an apparent inversion of the t^*/t formulation. But, the t and t^* did not represent performance times but WTPT scores collected under the Air Force's

JPMS Project. The variable t was an individual's WTPT score while t^* was the highest obtained score for the sample. Since the t^* and t values represented scores in which higher is better (the opposite of performance time), Faneuff and others' formulation is essentially equivalent in terms of interpretation, as the Carpenter and others t^*/t formulation.

2.7.1.2 The Formulation of Productive Capacity Under the Productive Capacity Project. AL/HRM proposed the PC formulation t/t^* (7). This is a simple inversion of the ratio proposed by Carpenter and others. This reformulation was made because of AL/HRM's concern that the original formulation of Carpenter and others does not result in a linear transformation of the estimated performance time data. This was perceived as a nuisance factor for the type of analyses AL/HRM was considering. A nonlinear transformation has the ability to adversely influence measures of linear relationship between two variables, such as the Pearson correlation coefficient.

The AL/HRM formulation does not have the desirable property of ranging from zero to one. Although PC scores under this formulation can range from one to ∞, the scores still maintain a degree of interpretability. A PC score of one means an individual is theoretically operating at maximum possible PC. Scores above one represent multiples of the fastest possible performance times. For instance, a PC score of two would imply that it takes the individual receiving this score takes twice as long as the fastest possible performance time, to complete the work.

2.7.1.3 The Formulation of Productive Capacity in Time Studies. Although time studies were not previously reviewed during the discussion of Air Force job performance measurement efforts, their methodology provides a possible PC formulation so they must now be reviewed. A time study is generally an Industrial Engineering technique used to derive time standards for completing certain job tasks and production-type jobs.

A first step in a time study is to clearly specify the operation to be studied. After the operation is clearly specified, a generally average worker, or operator, is selected to serve as the

subject of the study. The operator is then timed with a stopwatch by a qualified observer, for a specified number of cycles of the work. After the performance times are collected, the observer then assigns a performance rating reflective of the production rate of the operator.

The performance rating is used in "equitably determining the time required to perform a task by the normal operator after the observed values under study have been recorded" (29:325). In other words, the performance rating is used to adjust the time of the actual operator so that it reflects the time to be expected for a truly *normal* operator. If the selected operator worked faster than normal, as perceived by the observer, the observed time would be adjusted downward to reflect the normal time. Likewise, if the operator performed slower than normal, the observed time would be adjusted upward.

A common method of performance rating assumes that the normal operator is associated with a rating of 100, and performance greater than normal is indicated by values directly proportional to 100 (29:345). Thus, a rating of 120 would indicate that the operator's performance is 20% greater than normal, while a rating of 80 would indicate performance 20% below normal (29:345).

This time study performance rating can be interpreted as a PC measure for a given operator. The underlying formulation of the measure could be stated as $(t_{normal}/t) \times 100$, where t_{normal} is the time it would take a *normal* operator to do the task under study, and t is the time it takes the *actual* operator to complete the task.

This PC formulation offered a third option for standardizing the estimated performance time data collected under the Productive Capacity Project, provided the reasonable substitution of t_{avg} for t_{normal} is made. The quantity t_{avg}, the average time to complete the task, is virtually synonymous with t_{normal} and could be computed given the available Productive Capacity Project data.

2.7.2 Selecting a Task Weighting Scheme. Applying task weightings would give the tasks different levels of influence on the computed overall PC measure. This weighting is essential

if one is to allow *more important* tasks to have greater impact on overall PC. The question is which tasks are *more important*?

Tasks are known to differ on many dimensions such as criticality, time to complete them, learning difficulty, percent of the incumbents performing them, and percent of time incumbents spend on the tasks (40). Any of such factors could serve as a weighting factor, depending on the nature of the overall PC measure being computed.

In developing the WTPT, task clusters were weighted by the product of the mean recommended training emphasis rating and the cumulative percent time spent performing tasks in a cluster (23:6). The weights were used in determining how many tasks from each cluster to include in the WTPT. This weighting factor assigned weights (importance) to tasks based on how important the tasks were perceived in the training community and how much time airmen spend doing them. This appeared to be a reasonable weighting factor for selecting tasks for the WTPT, but did not appear so for computing overall PC measures. Since PC is a quantity-based measure of a worker's capacity to produce, it did not seem appropriate to let the training emphasis play a part in the weighting scheme since this did not seem to be an influencing factor on how much an airman can produce.

Carpenter and others, in the initial PC research, used a weighting scheme to weight the estimated performance times of individual's on the 10 task clusters when computing overall PC (5). But, it is not stated what the weighting scheme was.

2.7.3 Aggregating the Task-Level Data into an Overall Productive Capacity Measure. As just mentioned, in the initial PC research, Carpenter and others used a weighted average of the estimated performance times for the task clusters to compute an overall observed PC measure (5:23). But there was no mention of what the weighting scheme was. Unfortunately, this was the only research documented by the Air Force where job performance data were collected at the task or task cluster level and so required aggregation. The literature thus indicates that the only way

task-level data has been aggregated into overall PC measures was through weighted averaging of the task-level data.

2.7.4 Developing Prediction Models. There were three primary studies which involved the modeling of PC or performance time as the response, and the use of aptitude and experience as predictors. These were the studies by Carpenter and others, Faneuff and others and AL/HRM (5) (13) (38). The results of the studies were varied. Carpenter and others reported the highest R^2s of any of the studies using a two-predictor, first-order logistic model (5). Faneuff and others found that a first-order linear model fit their PC data better than a logistic model (13). Finally, AL/HRM found relatively good fit to untransformed time data using learning curve models (38).

2.8 Research Direction.

The reviewed literature provided some definite direction for the current research. First, The literature suggested four possibilities for meeting the first research objective, formulating a PC measure from task-level time data:

1. t^*/t
2. t/t^*
3. $(\frac{t_{evt}}{t}) \times 100$
4. t

Previous research offered only limited insight into how to meet the second research objective, selecting a task weighting scheme. In developing the WTPT, tasks were weighted by the mean recommended training emphasis rating and the cumulative percent time spent performing tasks in a cluster (23:6). However, such a weighting scheme did not appear appropriate for the current research because of the nature of the PC measure. (PC is a quantity-based measure of a worker's capacity to produce, and to weight it by mean recommended training emphasis rating and the

cumulative percent time spent did not appear to make sense.) This was the only weighting scheme discussed in the literature. Since it did not seem appropriate for the current research, the literature thus provided no particular direction for the second research objective.

The literature likewise offered only limited direction for the third objective, aggregating the task-level PC data. The only aggregation method discussed in the literature was weighted averaging of the task cluster-level data (5:6). However, this seemed to be a reasonable aggregation method and was chosen as the method of aggregation for this thesis.

The literature did provide significant guidance for the last and most important objective, developing prediction models. Three models having relevance to the current research (response of PC or performance time) were discussed in the literature. These models were:

1. Logistic model for predicting PC
2. Linear model for predicting PC
3. Learning curve model for predicting performance time

Since there was little or no guidance provided for the second and third research objectives, the research direction suggested by the literature can best be summarized in Figure 9. In Figure 9, the individual boxes indicate the response formulation and model type combinations which existed, given previous studies. The darkened boxes indicate which combinations were inappropriate due to response formulation and model type incompatibility. Written in the appropriate boxes are the studies that were done for a given response formulation and model type combination. An empty box indicates no studies have been done for a particular combination.

It was decided that this thesis would incorporate one of the PC formulation and model type combinations for which a previous study had been done. This was to take advantage of the information available as a result of the previous study. This left three choices:

1. Logistic Model with the t^*/t formulation

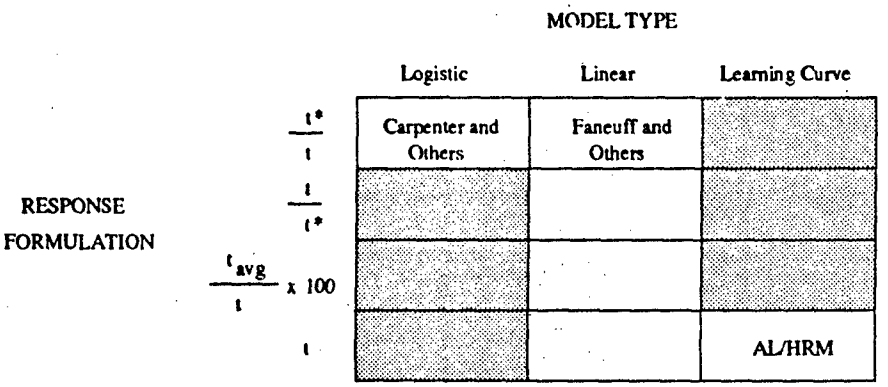


Figure 9. Graphical Representation of the Research Direction Suggested by the Literature

2. Linear Model with the t^*/t formulation

3. Learning curve model with t formulation

Because Carpenter and others, using the logistic model with the t^*/t formulation, obtained higher R^2s than Faneuff and others did with the linear model, the first combination above was determined a better alternative than the second. And, because the learning curve model seemed inappropriate for modeling overall job performance, the first combination also appeared better than the third. It was thus decided that given the estimated time data collected under the Productive Capacity Project, the response, PC, would be formulated as t^*/t, and the regression model for predicting it would take the form of the logistic model. The remainder of this thesis documents the research performed to develop the regression-based job performance model, using the PC formulation t^*/t and the logistic regression model.

III. Methodology

The last two chapters were designed to provide the reader with a substantial background on modeling, and the Air Force's job performance measurement R&D leading up to and including the first research effort under the Productive Capacity Project (21). This chapter describes the steps taken in developing the experimental mathematical models for predicting the job performance of Air Force Aerospace Ground Equipment (AGE) personnel given the estimated task performance time, aptitude and experience data collected under the Productive Capacity Project. Development of such descriptive models was of course the primary research objective of this thesis. This chapter begins with a brief overview of the subjects and data used to meet the research objectives. Following the overview of subjects and data, the specific steps taken to meet each research objective are discussed.

3.1 Subjects.

The experimental subjects were 204 airmen and NCOs studied by Leighton and others under the Air Force's Productive Capacity Project (21). The subjects were assigned to Air Force specialty 454X1, AGE. AGE personnel are the airmen responsible for inspecting, maintaining and repairing necessary ground equipment used to support aircraft and Ground Launched Cruise Missile (GLM) systems (8). Such ground equipment is called aerospace ground equipment and includes items such as electrical generators, heaters, hydraulic bomblifts, and air compressors.

The subjects were from the Air Force bases listed in Table 5. The procedures used to select the experimental subjects are described briefly in section 2.3.2.5 and in depth in the technical paper by Leighton and others (21). Figure 10 through Figure 13 describe some notable sample characteristics.

As can be seen in Figure 10, the vast majority of the sample were E-3 (Senior Airmen) or E-4s (Sergeants). Also, Figure 11 shows that most of the sample was from the 5 skill level, with

Grade	Frequency
E-2	10
E-3	55
E-4	119
E-5	16
Unknown	4

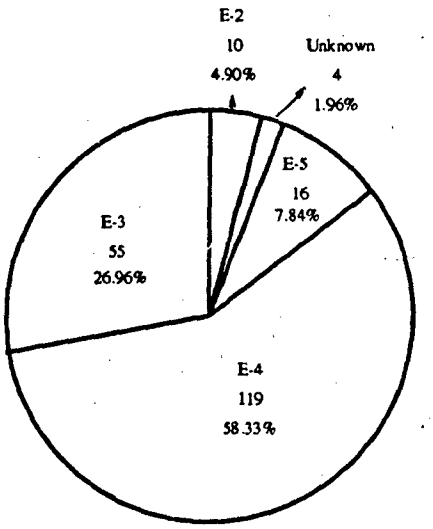


Figure 10. Frequency Distribution and Pie Chart of Subject Grade

Skill Level	Frequency
3	20
5	102
7	10
Unknown	72

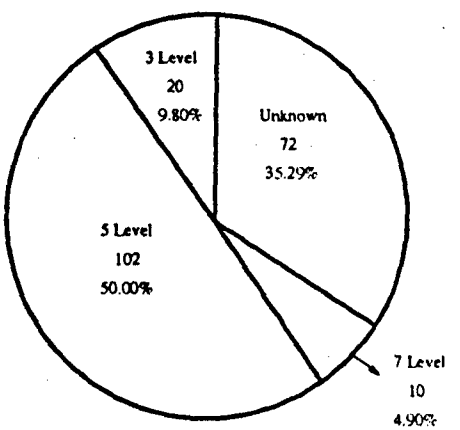


Figure 11. Frequency Distribution and Pie Chart of Subject Skill Level

Job Experience (Months)	Frequency
0-12	13
13-24	31
25-36	44
37-48	33
49-60	30
61-72	12
73-84	13
85-96	5
97-108	1
109-120	5
121-132	2
133-144	0
145-156	0
157-168	3
169-180	1
> 180	1
Unknown	10

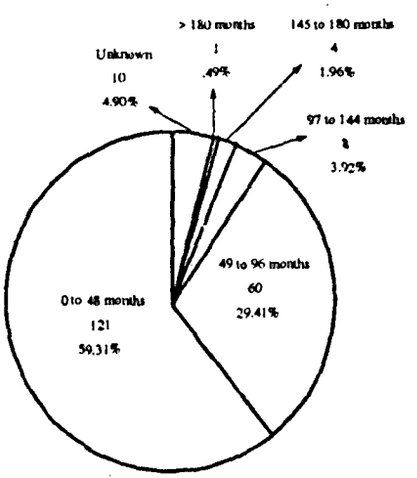


Figure 12. Frequency Distribution and Pie Chart of Subject Job Experience

ASVAB Mechanical Percentile Score	Frequency
46-55	12
56-65	40
66-75	54
76-85	34
86-95	31
96-99	10
Unknown	23

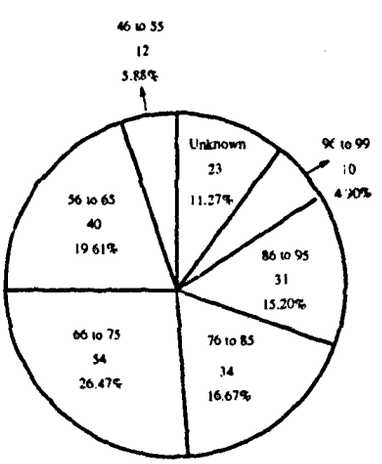


Figure 13. Frequency Distribution and Pie Chart of Subject Aptitude

much smaller numbers being from the 3 and 7 levels. In reference to the subjects' job experience, Figure 12 shows the majority of the sample were likely first-term airmen (121 our of 204, or 59.31%) as indicated by job experience between zero and 48 months (Some retrainees might also show a job experience less than or equal to 48 months but have significantly more Air Force experience.). Note also that 88.73% of the sample had eight years or less experience, and only about 2.45% had more than eight years experience. Also, observe that the experience data appear positively skewed, meaning the preponderance of subjects were associated with job experience measures near the low end of the experience range.

The aptitude distribution in Figure 13 indicates that the sample Mechanical aptitude percentile scores were distributed between 46 and 99. The ASVAB Mechanical score (M) distribution was considered important because the M score was used as the aptitude predictor in the regression modeling. The M score was selected as the aptitude predictor because there was a minimum M score requirement for entering the AGE specialty. This indicated that mechanical aptitude had been previously identified by the Air Force as being somehow related to performance in the AGE specialty, thus M seemed appropriate as an aptitude predictor variable for the current study. The Mechanical aptitude score distribution was restricted to values generally greater than 51 because this is the current minimum Mechanical aptitude score requirement for entering the job. (An additional requirement of an Electronic percentile score (E) of at least 33 also exists for entering the job (8).)

Table 9 provides a two-frequency distribution of subject aptitude by job experience. Recall that aptitude and experience are the predictor variables. The two-way frequency distribution was provided to offer insight as to what the true effective range of the estimated regression model is. In other words, sparse or null cells in regions of Table 9 indicate that the regression model should be interpreted cautiously in such regions. This is because the shape of the estimated response surface in such areas was determined by relatively few data points. Note that the matrix depicted

Table 9. Two-Way Frequency Distribution of Sample Aptitude by Job Experience

Months of Job Experience	ASVAB Mechanical Percentile Score							
	46–55	56–65	66–75	76–85	86–95	96–99	Unknown	Total
0–12	0	5	1	4	3	0	0	13
13–24	4	5	12	5	2	2	1	31
25–36	1	10	12	9	8	4	0	44
37–48	2	8	10	6	5	1	1	33
49–60	1	6	8	7	3	2	3	30
61–72	0	0	3	1	2	1	5	12
73–84	2	2	0	1	5	0	3	13
85–96	0	1	2	1	0	0	1	5
97–108	1	0	0	0	0	0	0	1
109–120	0	0	0	0	1	0	4	5
121–132	0	0	1	0	0	0	1	2
133–144	0	0	0	0	0	0	0	0
145–156	0	0	0	0	0	0	0	0
157–168	0	0	0	0	1	0	2	3
169–180	0	0	0	0	0	0	1	1
> 180	0	1	0	0	0	0	0	1
Unknown	1	2	5	0	1	0	1	10
Total	12	40	54	34	31	10	23	204

in Table 9 is very sparse beyond 96 months of job experience. The estimated models may thus be tenuous in that region.

In summary, the sample tended to be E-3s and E-4s, with skill levels around 5. Further, the airmen tended to have less than eight years of job experience, and aptitude covering the somewhat restricted range of 46 to 99.

3.2 Data.

As previously mentioned, the data used in this thesis were collected under the Air Force's Productive Capacity Project, by Leighton and others, between March and September 1990 (21). A brief overview of the Leighton and others' research, to include data collection, was included in Section 2.3.2.5. Again, the reader is referred to (21) for a complete description of that research.

The primary data used were the estimated task performance times provided by each subject's supervisor. Associated with each subject was his or her supervisor's estimates of how fast he or she could complete each of 50 job tasks while simultaneously working as quickly as possible and maintaining an acceptable level of task quality. Complete task descriptions are included in Table 22 at Appendix A.

The tasks selected for analysis were those that tend to be performed by fairly junior and intermediate personnel. The tasks included mostly hands-on production-type tasks as opposed to the supervisory or management tasks that more senior personnel perform. With this in mind, the sample described in the previous section appeared to be a fairly reasonable sample as indicated by the grade, skill level and experience characteristics provided.

Not all subjects had a complete set of 50 task ratings. Some supervisors did not provide all ratings for all subjects. As a result, a relatively small number of missing values existed.

As previously indicated, other primary data used for the analyses included the subjects' self-reported level of job experience, and the subjects' Mechanical composite score from the ASVAB obtained when applying for enlistment. These data were used as predictors in the mathematical prediction of the subjects' productive capacity. As previously mentioned, the M aptitude score was chosen as the aptitude predictor because scores on this composite help determine a recruit's eligibility for entering the AGE specialty.

Secondary data of interest were the subjects' Job Knowledge Test percent correct scores (JKT), the supervisors' global or overall estimates of the subjects' PC (GPC), and a PC measure derived from actual stopwatch times of a limited subsample of the subjects (MTPC). These measures were used as a basis for comparison for the regression model results derived in this thesis. Figure 14 provides a graphical representation of the data used in the analyses.

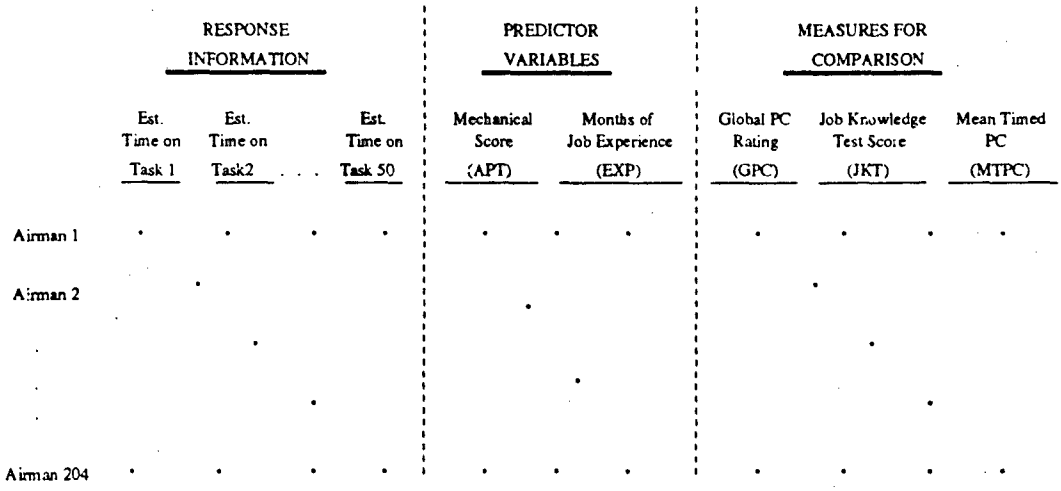


Figure 14. Graphical Representation of the Data Used in the Analyses

3.3 Procedure.

The preceding sections provided a brief overview of the experimental subject sample and the relevant data collected. Discussion may now proceed to the actual steps taken to meet the research objectives. The reader may wish to keep in mind that although the primary research objective was to develop regression-based job performance models, the first three research objectives (see Section 1.3) were concerned only with the response information, the estimated task performance times.

3.3.1 Formulating a Productive Capacity Measure from Estimated Task Performance Times.

As mentioned in Chapter 1, it was necessary to transform the estimated task performance times to give them interpretability and to allow them to be aggregated across tasks.

In reference to Figure 4, the formulation of a PC measure from the raw time data is associated with first phase of the model building process, data collection and preparation. Of course, the PC formulation was only concerned with the preparation part because the data had already been collected. In reference to Figure 14, the PC formulation involved editing and transforming the data under the *Est. Time on Task i* columns.

3.3.1.1 Defining Task-Level Productive Capacity. Task-level PC was defined according to the Carpenter and others formulation t^*/t (5:21). Recall that t^* is the fastest possible completion time for a given task, and t is a subject's completion time. In reference to Figure 14, the *ts* are the entries under the *Est. Time on Task i* columns. The t^*s were derived from the minimum observed time in each such column.

As explained at the end of the previous chapter, the t^*/t formulation was selected over the other possible formulations. The t^*/t formulation has some desirable characteristics which made it a reasonable choice. First, unlike the other formulations, that of Carpenter and others yields values that range between zero and one, thus lending themselves to logistic regression models. Recall that it was the logistic regression models of Carpenter and others that yielded the highest reported model R^2s for any of the PC studies (5) (13) (38). Second, the Carpenter and others formulation maintains the desirable property of being nicely interpretable. It can be interpreted as an individuals work capacity as a proportion of maximum possible capacity.

3.3.1.2 Editing the Raw Estimated Time Data. Before the PC measures were computed from the estimated task performance times, the estimated times were edited to control for serious outliers. As Neter, Wasserman and Kutner (1990) point out, "Outliers can cause great difficulty." (27:121) They describe how when least-squares estimation is used in trying to predict a response, a fitted surface can be pulled disproportionately toward an outlier. They suggest discarding an outlier "if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type circumstance." (27:122) The reason that editing was justified with the raw estimated time data is because the format in which the time estimates were collected was a type of *free-response* format. This means that there was no limitation on the answers that could be given. Recall that when the supervisors provided their time estimates, they were provided with previously created benchmark scales showing SMEs' opinions as to what the fastest, normal and slowest completion times were. However, these were to be used as *take-it-or-*

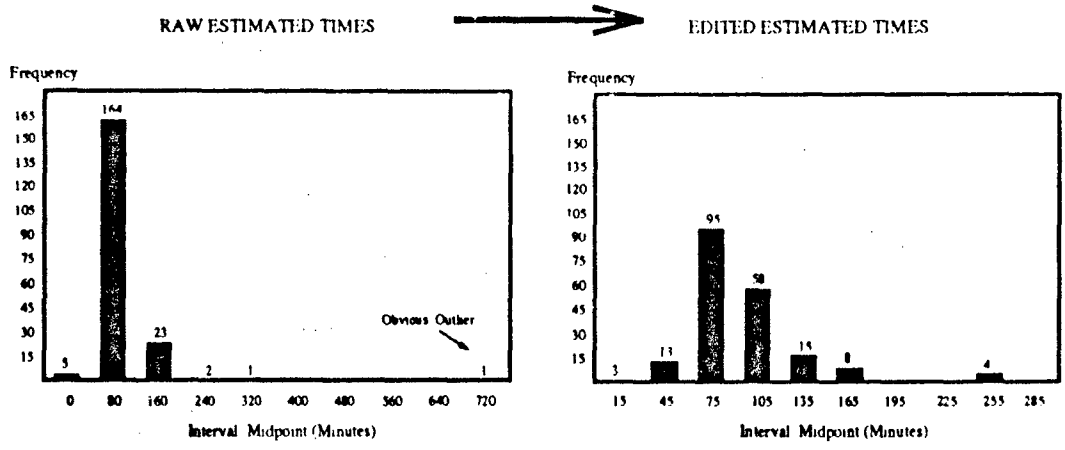


Figure 15. Histograms of the Raw and Edited Estimated Times for Task G179

leave-it guidance for the supervisors in making their estimates. The supervisors were not required to keep their estimates between the fast and slow time benchmark times if they did not want to. This resulted in a free-response format. One problem with the free response format is that responses have a tendency to widely vary, even to extremes. This necessitated the need for data editing.

To control for potential outliers, the raw time estimates for a given task were edited by *pulling in* all values that were beyond ± three standard deviations from mean. In other words, all values beyond three standard deviations were recoded to a value of the *mean ± three standard deviations*. This was done because for a distribution of measurements that is approximately bell-shaped, the interval between ± three standard deviations will contain almost all the measurements (25:9). Thus, anything beyond these limits was considered an outlier, and recoded. Values beyond three standard deviations from the mean were recoded and not discarded because these outlying data were not considered to be transcription errors or results of some other error, but an estimate from a supervisor who did not happen to agree with the range of times provided. The recoding was thus done to retain the information contained in the outlying points while keeping some consistency in the ratings and keeping the variance at a reasonable level. Figure 15 provides an example of the effects of the editing on the raw data for one task, G179.

3.3.1.3 Computing the Productive Capacity Measure at the Task Level. As previously mentioned, task-level PC was defined as t^*/t. In the actual computation, t^* for a task was computed as .99× (*minimum estimated time for the task (after editing)*). Again, in Figure 14, the *ts* are the entries under the *Est. Time on Task i* columns and t^*s are derived from the minimum estimated time in each such column. This computation of t^* accounted for the fact that true *fastest possible time* was probably not recorded for this sample, but is likely somewhat less than the sample minimum. After computing t^* in this fashion, the PC measure t^*/t was computed for each individual for each task.

3.3.1.4 Editing the Productive Capacity Measure at the Task Level. After computing the task-level PCs, a review of their histograms indicated that the editing of the raw estimated times was not enough to control for serious outliers. Several of the task PC distributions still indicated additional obvious outliers. This indicated the need for further editing.

The task-level PCs were edited much the same as the raw estimated times. For each task, PC measures beyond ± three standard deviations from the mean were *pulled in* to values of the *mean ± three standard deviations*. Unlike the editing of the raw estimated times, this editing influenced the interpretability of the PC measure. Recall that PC is interpreted as an individual's output as a proportion of maximum possible output. As an example of how the interpretability was influenced, consider an example where the *mean ± three standard deviations* defines the range of .2–.8. Assume that all values outside of this range are considered extreme outliers and recoded as .2 or .8, depending on which side of the interval they fall. The recoding is done because values outside of the range *mean ± three standard deviations* are considered impossible. After recoding, the range of PC values is not zero to one but .2 to .8. Since .2 represents the new *lowest possible* output level, it must correspond to a PC of zero. Likewise, since .8 represents the new *highest possible* output level, it must correspond to a PC of one. To make .2 and .8 correspond to zero and one respectively, the rescaling transformation in Equation 17 was made on the edited PC values for

the task. The rescaling ensured the interpretability of the PC measure as an individual's output as a proportion of maximum possible output.

The rescaling transformation function in Equation 17 is a linear function of the original PC data. This means that the transformed data will exhibit exactly the same linear associations (same correlation coefficient, same linear regression results, etc.) with other variables as the untransformed data. The rescaling may, however, influence logistic regression results because the logistic model is not a linear model in its original form.

A small adjustment made to the rescaled values was to recode the rescaled value of zero to .01, and rescaled value of one to .99. This was to ensure that the logistic model would be defined for all computed rescaled values. (The range of the logistic function does not include zero or one.)

$$PC_{rescaled} = \frac{PC_{obs} - PC_{min}}{PC_{max} - PC_{min}} \tag{17}$$

where

$PC_{rescaled}$ = *PC rescaled to 0-1 space*

PC_{obs} = *Observed value of PC*

PC_{min} = *Minimum observed value of PC*

PC_{max} = *Maximum observed value of PC.*

After reviewing the histograms of the edited, rescaled PC values, 17 tasks still showed serious outliers. These were G171, G179, G181, H238, I251, I255, I264, I265, I283, I284, I299, J332, J347, L406, M444, N486, P549. One final editing and rescaling was applied. This time, outliers from the 17 tasks were identified through subjective judgement by the author. The outliers were then *pulled in* to the closest reasonable observed value. The reedited PCs were then rescaled according to Equation 17, and the adjustments to the zero and one values were made. This completed the

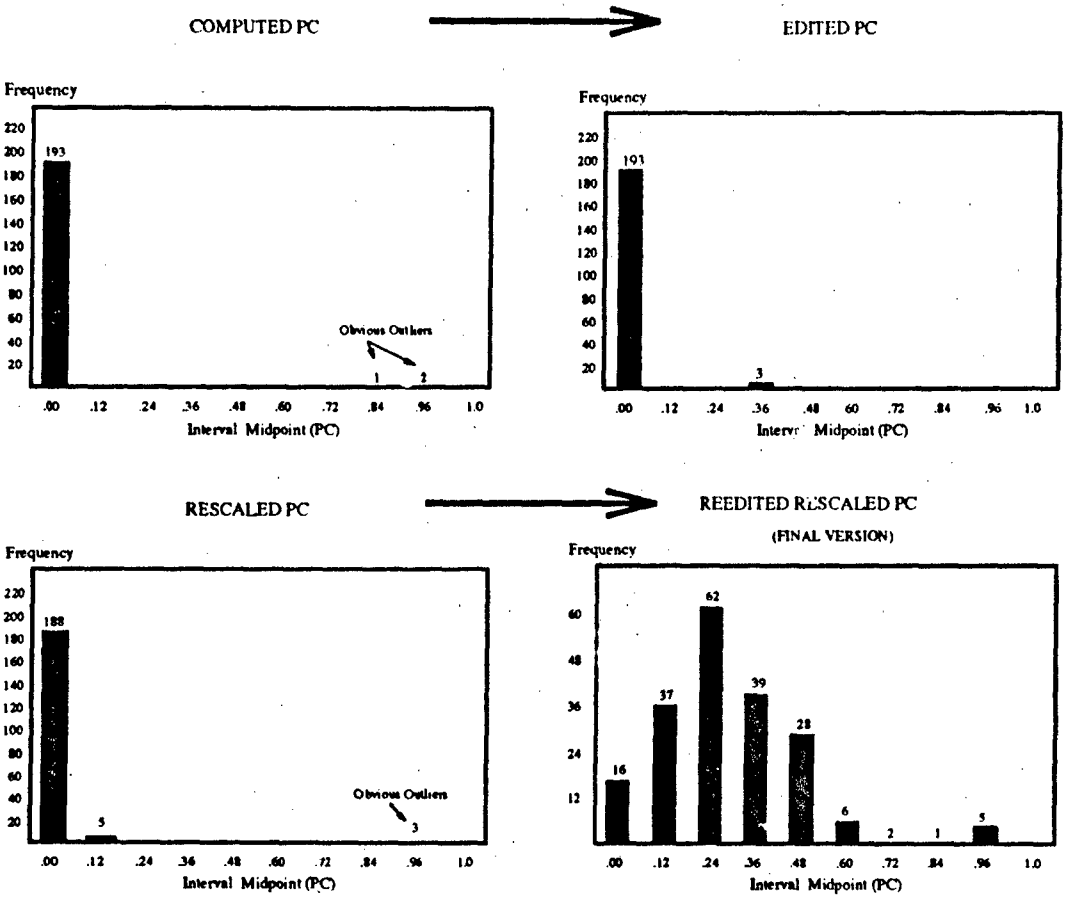


Figure 16. Histograms of the Productive Capacity Values in the Editing Process for Task G179

data editing for the PC task variables. Figure 16 provides an example to show the effects of the PC editing and rescaling for task G179.

After the final editing, summary statistics were computed for the task-level PCs for comparison to the summary statistics for the associated raw estimated times. This comparison was made primarily to determine if the editing had the desired effects of outlier and variance control. Of primary interest was the *coefficient of variation*. The coefficient of variation, CV, is a measure of the dispersion of the distribution of a variable. The computational formula for CV is shown in Equation 18 (26:388).

$$CV = \frac{s}{\bar{X}} \tag{18}$$

where

CV = *coefficient of variation*

s = *standard deviation*

$\bar{X}$ = *arithmetic mean.*

CV expresses a distribution's dispersion relative to the distribution mean, thus making the measure comparable across variables with different distributions. A CV of less than one is generally indicative that a distribution is not highly variable, thus partially indicating that the distribution is not subject to severe outliers.

Having computed CV for both the raw estimated times and final PC measures for each task, it was possible to assess the effects of the t^*/t transformation, the editing of the raw times, and the editing of the PCs on the response data.

3.3.2 Selecting a Task Weighting Scheme. The selection and application of a task weighting scheme still involved the data preparation phase of the regression model building process depicted in Figure 4. In reference to Figure 14, the selection of a weighting scheme involved the identification of appropriate weights for each *Est. Time on Task i* column to give the data derived from each column an assigned level of importance. This was to give the task-level data varying levels of influence when computing an overall measure.

Because the PC measure is time-based and reflective of overall worker output, it seemed most appropriate to weight the tasks by the average amount of time individuals spend doing each task. If the individual under study is slow on some tasks and fast on others, it is necessary to consider the relative amount of time spent on each task to accurately assess overall capacity. To illustrate,

consider the extreme situation in which a worker is exceptionally fast on all but one job task. And, say the worker is exceptionally slow on that one outstanding task. If the job requires the individual to perform the outstanding task 99% of the time, his or her productive capacity should be comparatively low. This is despite the fact that his or her performance is exceptionally good on the other numerous tasks that are infrequently performed.

The Occupational Measurement Squadron collects relative performance time data as part of their periodic surveys of the AFSs (40). One such measure outlined in the Occupational Survey Report is *Average Percent Time Spent Performing Duties* (40:23). In the report, the data is broken out by skill level. The task weightings used in this thesis were computed as an average of the *average percent time spent* for the skill levels that would generally be expected to do the types of hands-on tasks under study (skill levels 3, 5 and 7). Because of the nature of the available *average percent time spent* data, weights had to be derived for each duty area, and the duty area weight was applied to each task from that duty area.

Overall, the selected weighting scheme was designed to give greatest importance to tasks from the duty areas that are performed most often by 3, 5 and 7 skill level airmen.

3.3.3 Aggregating the Task-Level Data into an Overall Productive Capacity Measure. As with the first two research objectives, this one dealt with the data preparation phase of the model building process depicted in Figure 4.

Having computed the task weights, it was possible to define and compute aggregate or overall PC per individual. The following discussion describes how this was done.

3.3.3.1 Defining and Computing Overall Productive Capacity. To derive a single PC measure for an individual from his or her task-level data, it was necessary to somehow collapse task-level ratings into a single overall measure. Figure 17 presents a graphical illustration of the task-level data aggregation.

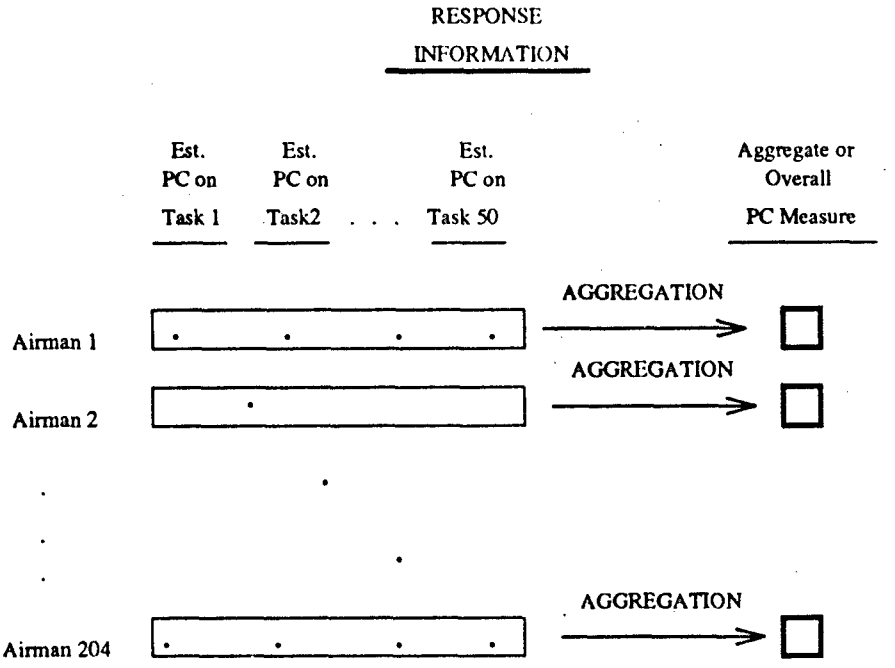


Figure 17. Graphical Representation of the Task-Level Data Aggregation

This aggregation of the data was accomplished through weighted averaging. *Aggregate* or *overall* PC, then, was defined as a weighted average of the subjects' final edited and rescaled task-level PCs. Weighted averaging was used because previous studies had successfully used weighted averaging as an aggregation method (5). Also, weighted averaging is a commonly accepted and frequently applied statistical technique used to aggregate data (of the same units) that differ on known dimensions. Equation 19 shows a mathematical representation of how the aggregate PC measures were defined (21).

$$PC_{wavg} = \frac{\sum_{i=1}^{n} w_i PC_i}{\sum_{i=1}^{n} w_i} \tag{19}$$

where

PC_{wavg} = *a weighted average of task-level PCs*

PC_i = *the individual's PC on task i*

w_i = *the weight for task i*

n = *the number of task measurements for the individual.*

A simple, or unweighted, average was also computed for strictly comparative purposes. The unweighted and weighted average PC values were compared through summary statistics and correlational analyses. The correlation statistic used was the *Pearson product-moment correlation coefficient*, r (26:429). The computation for r is shown in Equation 20.

$$r = \frac{n\sum_{i=1}^{n} U_i Y_i - (\sum_{i=1}^{n} U_i)(\sum_{i=1}^{n} V_i)}{\sqrt{[n\sum_{i=1}^{n} U_i{}^2 - (\sum_{i=1}^{n} U_i)^2][n\sum_{i=1}^{n} V_i{}^2 - (\sum_{i=1}^{n} V_i)^2]}} \tag{20}$$

where

r = *Pearson product-moment correlation coefficient*

i = *observation number*

n = *number of observations of U and V*

U_i = *observation i of a variable U*

V_i = *observation i of a variable V.*

The Pearson correlation coefficient is a measure of linear association between two variables. The coefficient ranges between -1.0 and 1.0. Measures near -1.0 and 1.0 indicate a high degree of linear relationship. A negativ oefficient means the measures are inversely related, or one measure tends to be high when the other is low.

The unweighted and weighted average PCs were compared via summary statistics and r to determine if the measures were unique. The idea was that if the weighted and unweighted measures were statistically similar and highly positively correlated, then the weighting added no uniqueness

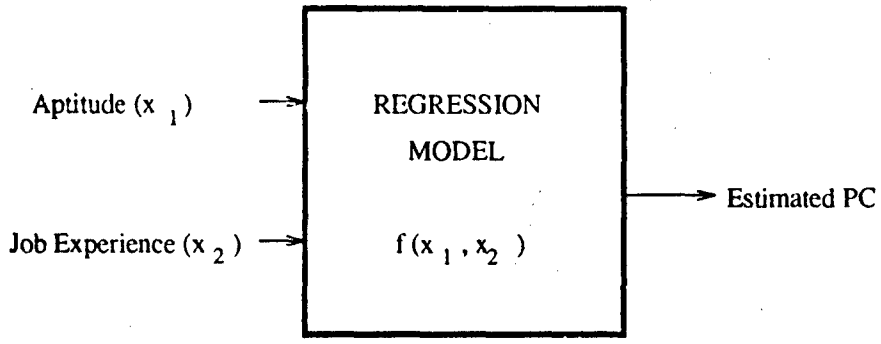


Figure 18. Graphical Representation of the Regression Models Developed

to the overall measure. Similar summary statistics and high correlation would thus indicate that the weighting scheme added nothing to the computation of overall PC beyond what could be gained by simple averaging.

3.3.4 Developing Prediction Models. After computing the aggregate PC variables, it was possible to begin the modeling phase. The following sections describe the steps taken to complete the regression modeling at the aggregate level and also at the task-level. Figure 18 provides a graphical representation of the regression models to be developed. The goal of the regression analysis was to determine the β parameter estimates that would define the mathematical function of the predictors depicted in the model box.

3.3.4.1 Editing the Predictor Variables. The previously discussed research objectives each dealt with preparation of the response data for the regression modeling of job performance. Like the response data, the predictor data had to be prepared in accordance with the first phase of the model building process depicted in Figure 4. In reference to Figure 14, the graphical data file depiction, the following editing procedures were applied to the columns under the heading *Predictor Variables*.

As with outlying response values, outlying predictor values can be problematic. "Outlying cases may involve large residuals and often have dramatic effects on the fitted least-squares re-

gression function." (27:392) Recall that the predictor variables are aptitude (ASVAB Mechanical percentile scores) and experience (months of job experience). Frequency distributions and pie charts for these variables were provided in Table 9, Figure 12 and Figure 13.

The frequency distribution of aptitude scores indicated that there were no obvious outliers or other apparent problems with the aptitude data. The scores appeared near normally distributed between 46 and 99. The experience variable's distribution appeared positively skewed with the vast majority of the observations (88.73%) having 96 or less months of job experience. Note that the frequency distribution shows one potential outlier with a value greater than 180 months. The actual value recorded for this observation was 283 months, well beyond the next highest value of 169. A review of the data file showed that no subject had more than 195 months of total Air Force experience. It is of course impossible to have more Air Force job experience than overall Air Force experience thus the value of 283 was identified as a miscoding. The case was dropped from further analyses.

3.3.4.2 Fitting the Regression Models. The editing of the predictor variables concluded the data preparation phase of the model building process. The next phases, according to Figure 4, were reduction of the number of predictor variables and model refinement and selection. The following discussion describes these phases applied to the current research.

Recall from the literature review that the model which yielded the highest R^2s among the Air Force's PC studies was the logistic model used by Carpenter and others (5:21) (13) (38). With this result in mind, a logistic model was fit to the PC data for each of the 50 tasks, and also to the weighted and unweighted average PCs. The logistic model and logistic regression were discussed only briefly in the previous chapter. Following is a more in-depth discussion.

The logistic regression model is a model that is frequently applied in situations where the response variable is binary, zero or one. In such situations, the observations are often classified into groups based on values of one or more predictor variables. Thus, grouping of observations allows

the individual zero/one response observations to be collapsed into a proportion for the group. The zero/one response often indicates an observation's possession (one) or lack (zero) of some trait of interest. Grouping observations collapses the responses into a single measure representing the proportion of observations possessing the trait. The logistic function, being restricted to the range between zero and one, is ideally suited for modeling such proportions given known levels of the predictors. Logistic regression is thus frequently used to predict the proportion of individuals in a given group which possess the trait of interest.

A general form of the logistic model is expressed in Equation 21 (17:25–26).

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} + \epsilon \tag{21}$$

where

$\pi(x)$ = *a response variable (ranging from 0 to 1)*

$g(x)$ = *some function of the predictor variables (linear in the β parameters)*

ϵ = *the model error terms.*

Note that the logistic model is not a linear model because it is not linear in the β parameters which would be contained in the function $g(x)$ (The function $g(x)$ is linear however. This fact will be used later.)

The logistic function is generally S-shaped as depicted in Figure 19 and Figure 20. These represent example plots of logistic functions with one and two predictors, respectively. The addition of higher-order and interaction terms and the nature of the relationship between the variables can cause the logistic function to take on shapes other than the standard S-shape. This will be shown in Chapter 4.

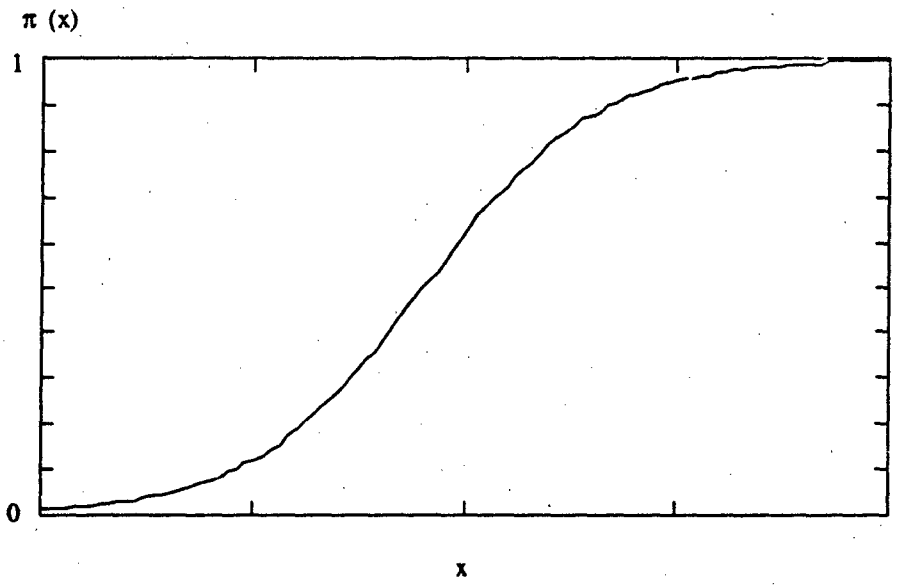


Figure 19. Plot of a First-Order Logistic Function with a Single Predictor

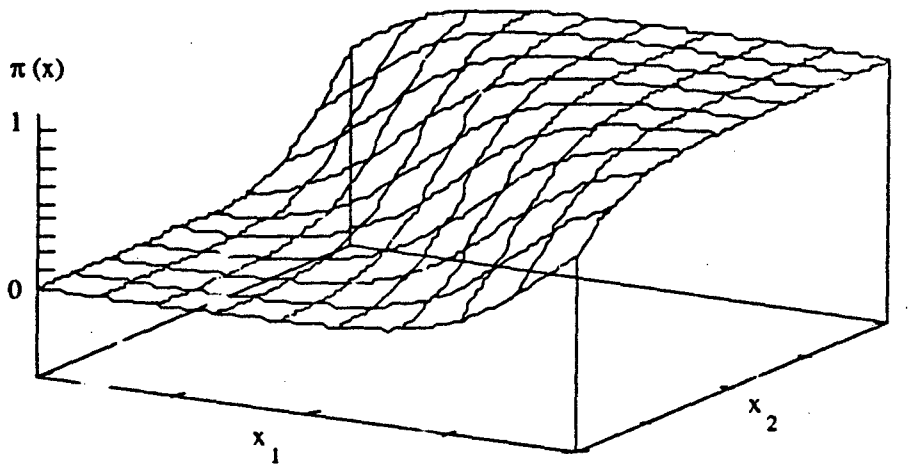


Figure 20. Plot of a First-Order Logistic Function with Two Predictors

As mentioned, the logistic model is not linear in its original form. But, it can be linearized using the *logit* transformation. The logit transformation is shown in Equation 22.

$$Logit\ of\ \pi(x) = ln(\frac{\pi(x)}{1-\pi(x)}) \tag{22}$$

Transforming the response (a proportion) through the logit transformation allows the logistic function to be written in linear form as in Equation 23. The linearized logistic function is called the *logit response function* (28:583).

$$ln(\frac{\pi(x)}{1-\pi(x)}) = g(x) + \epsilon \tag{23}$$

where

$\pi(x)$	=	*a response variable (ranging from 0 to 1)*
$ln(\frac{\pi(x)}{1-\pi(x)})$	=	*the logit of the response,* $\pi(x)$
$g(x)$	=	*some function of the predictor variables (linear in the coefficients)*
ϵ	=	*the model error terms.*

Although the logistic model can be expressed as a linear model, standard linear regression cannot be applied if the response data are originally binary and the analyst wishes to apply standard linear regression inferential statistics. Recall from the linear regression discussion in Section 2.1.1 that application of the linear regression inferential statistics requires the model assumption that the error terms, ϵ, are distributed $N(0, \sigma^2)$. It so happens when the response data are originally binary, the error terms are not normally distributed, but binomially distributed (17:7). Also, there is nonconstant error variance (*heteroscedasticity*) across varying levels of the predictors (28:581). These facts indicate that ordinary least squares estimation of the model parameters is inappropriate. When there are a sufficient number of repeat observations at each level of the predictors, the

parameter estimates can be obtained via weighted least squares (28:584-589). Otherwise, the parameters may be estimated using maximum likelihood estimation (28:589-595).

As a review, logistic regression is often applied when the response data are binary. And, the logistic regression model is characterized by three properties:

1. Nonnormal error terms
2. Nonconstant error variance
3. A constrained response function (between zero and one) (28:580-581)

These properties make the use of ordinary least squares estimation of the parameters inappropriate.

The above discussion of logistic regression assumes that the response data are originally binary, zero or one, data. If the response data are proportions, but not derived from binary data, an adaption of logistic regression is possible (see Reference (5)). Productive capacity, formulated as t^*/t, is one such proportion which may be modeled with the adaption of the logistic regression model. When the proportional response data are not derived from binary data, the logistic regression model is not necessarily characterized by nonnormal error terms and nonconstant error variance. This means that estimation of model parameters through ordinary least squares estimation may be possible. There is of course the requirement to check the linear regression model assumptions. Thus, the adaption of the logistic regression model to the nonbinary response case involves:

1. Use of the logistic function
2. Linearization of the logistic function through creation of the logit response function
3. Estimation of the model parameters using ordinary least squares
4. Aptness analysis to check normality of error terms

This adapted logistic regression model was used to model PC in this thesis.

Although primary interest was in predicting the aggregate or overall PC measures, the task-level regressions were run as a screening exercise to identify any trends in the relationships between the predictors and PC across tasks. This was to provide insight as to whether the number of predictor terms might be reduced (the second phase of the model building process). The adapted logistic model that was fit was a full second-order model to include aptitude/experience interaction terms. A full second-order model was selected to account for any curvature or interaction effects that may not have been accounted for with a first-order model. The model that was fit at the task and aggregate level can be found in Equation 24.

$$PC = \frac{e^{\beta_0+\beta_1 x_1^2+\beta_2 x_2^2+\beta_3 x_1 x_2+\beta_4 x_1+\beta_5 x_2}}{1+e^{\beta_0+\beta_1 x_1^2+\beta_2 x_2^2+\beta_3 x_1 x_2+\beta_4 x_1+\beta_5 x_2}} + \epsilon \tag{24}$$

where

PC	=	*productive capacity*
x_1	=	*ASVAB Mechanical percentile score*
x_2	=	*months of job experience*
$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$	=	*parameters to be estimated*
ϵ	=	*model error terms.*

The logistic model in Equation 24 was written as the linear model in Equation 25. Writing the equation in this fashion (linear in the parameters) allowed the model parameters to be estimated using least-squares regression.

$$ln(\frac{PC}{1-PC}) = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_1 x_2 + \beta_4 x_1 + \beta_5 x_2 + \epsilon \tag{25}$$

where

$ln(\frac{PC}{1-PC})$	=	*the logit of productive capacity*
x_1	=	*ASVAB Mechanical percentile score*
x_2	=	*months of job experience*
$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$	=	*parameters to be estimated*
ϵ	=	*the model error terms.*

After the logistic models were fit to the 50 tasks and the aggregate measures, a forward stepwise regression was run for the aggregate case (weighted average PC). This was in accordance with the second phase of the model building process, reducing the number of predictor variables.

After performing the stepwise regression, the resulting aggregate model was subjected to an aptness analysis to include a plot of residuals vs. predicted values and a normal probability plot. In reference to Figure 4, the aptness analysis concerns the third phase of the model building process, model refinement and selection.

After completing the regression and aptness analyses, predicted PC values were obtained for the aggregate model for use in subsequent correlational analyses. Recall that the logit response function in Equation 25 yields predicted values not for PC, but for the logit of PC. As a result, predicted PC measures were derived from the predicted logits using Equation 26.

$$\widehat{PC} = \frac{e^{\widehat{PC}_{logit}}}{1 + e^{\widehat{PC}_{logit}}} \tag{26}$$

where

$\widehat{PC}$	=	*predicted productive capacity*
$\widehat{PC}_{logit}$	=	*the predicted logit of productive capacity.*

3.3.4.3 Analysis of the Aggregate Measures. Once the predicted aggregate PC measures were computed, they were correlated with other measures of job performance to include JKT (Job Knowledge Test) scores, GPC (supervisor's global PC ratings), and MTPC (mean timed PC). The measure of correlation was r, the Pearson correlation coefficient.

The variable MTPC was created by AL/HRM using the PC formulation t/t^*, an inversion of the formulation of Carpenter and others (21) (7). MTPC was computed as the average PC across a limited number of tasks where the t values where derived through actual timing of tasks as opposed to supervisor estimation. Because MTPC was computed from t/t^* values, higher values indicated lower performance levels. This means that a negative Pearson correlation coefficient would be expected between MTPC, and a variable whose higher values indicate better performance.

This correlational analysis was to provide insight as to whether the aggregation method, weighting scheme and fitted model were effective in capturing an individual's true overall PC. Significant correlation with other performance measures was to be interpreted as evidence that the aggregation method, weighting scheme and fitted model were appropriate.

Last, the fitted logistic response surfaces were plotted for the weighted aggregate variable to provide a graphic illustration of the fitted model. Surfaces were plotted for the entire effective range of the predictor variables. Finally, response surfaces rescaled to zero/one space (see Equation 17) were also plotted to increase the interpretability of the plots.

IV. Results

The preceding chapter specified in detail exactly what steps were taken to meet each research objective. This chapter discusses the results of applying those steps and offers further discussions on the significance of the research findings.

4.1 Formulation of a Productive Capacity Measure from Estimated Task Performance Times.

As previously indicated, the primary response data used in the analyses were the supervisors' estimates of the subjects' task completion times. In their raw form, the time estimates tended to widely vary within a task, sometimes covering an unbelievable range of values. This implied the need for editing of the raw time data to control for serious outlying cases. Table 10 provides summary statistics for the raw time estimates, illustrating the sometimes extreme variation for a task. For instance, note task G171 which shows an extremely wide range of values for the raw estimated times. The raw times ranged from a minimum of one to a maximum of 2880 minutes. It was considered highly unlikely that the true range of times is so variable. This led to the editing as described in Section 3.3.1.2.

After the edited estimated times were computed, the task-level PCs were computed using the t^*/t formulation of Carpenter and others (5:21). These required further editing and rescaling as described in Section 3.3.1.4. Table 11 shows the summary statistics for the final edited and rescaled task PCs.

For the final edited values of the task PCs, the means ranged from .12 to .56 across tasks. The standard deviations for the tasks ranged between .12 and .22. Note, in particular, that in only two cases was the coefficient of variation, CV, noticeably greater than one (for task 1299 and J332). This is a very general indication that the task-level PC data are not highly dispersed relative to the task means, and thus are probably not highly influenced by extreme outliers. In contrast, Table 10

Table 10. Summary Statistics for the Raw Estimated Task Performance Times (in Minutes)

Task	n	Mean	S.D.	Minimum	Maximum	CV
E120	201	11.38	5.16	2.00	35.00	.45
E143	201	9.70	3.90	2.00	30.00	.40
F153	194	17.43	19.60	5.00	270.00	1.12
F154	201	17.17	9.05	5.00	120.00	.53
F155	200	15.06	6.62	3.00	60.00	.44
F157	200	17.59	8.89	5.00	90.00	.51
F162	200	24.61	16.91	4.00	240.00	.69
G171	200	82.88	215.76	1.00	2880.00	2.60
G179	196	89.36	57.81	1.25	720.00	.65
G181	199	274.84	341.74	2.50	2880.00	1.24
H202	201	14.43	126.60	2.00	1800.00	8.77
H203	200	27.25	66.83	10.00	720.00	2.45
H209	198	33.51	12.05	15.00	90.00	.36
H215	200	20.32	9.60	10.00	90.00	.47
H236	200	12.89	11.35	2.00	120.00	.88
H237	200	10.37	6.31	3.00	60.00	.61
H238	199	25.87	21.92	.60	300.00	.85
I247	200	10.31	6.06	1.00	60.00	.59
I248	200	13.29	12.92	3.00	180.00	.97
I251	195	50.94	18.06	1.00	150.00	.35
I255	199	154.62	54.22	3.00	480.00	.35
I260	201	24.47	62.98	8.00	900.00	2.57
I264	200	70.57	71.74	1.00	720.00	1.02
I275	200	67.49	213.94	1.00	3060.00	3.17
I283	201	63.82	189.14	1.00	2700.00	2.96
I284	200	21.39	11.77	.53	120.00	.55
I286	200	57.94	27.71	20.00	280.00	.48
I299	201	121.24	49.39	1.92	480.00	.41
I300	201	19.93	7.29	5.00	60.00	.37
J332	198	142.85	209.05	2.75	2880.00	1.46
J340	198	33.78	12.91	3.00	120.00	.38
J347	198	65.07	55.38	1.00	480.00	.85
J355	198	42.42	19.18	15.00	165.00	.45
L406	200	36.80	34.64	1.00	480.00	.94
L421	199	28.99	12.88	10.00	120.00	.44
L436	196	32.27	25.55	10.00	285.00	.79
L437	200	10.28	6.72	2.00	60.00	.65
M444	196	65.10	39.21	1.00	480.00	.60
M446	200	63.68	201.76	20.00	2880.00	3.17
M447	195	20.19	14.08	5.00	180.00	.70
N475	184	20.63	8.31	5.00	60.00	.40
N477	201	37.12	13.14	15.00	120.00	.35
N486	201	80.40	27.63	1.00	240.00	.34
N487	201	40.34	105.26	10.00	1500.00	2.61
N488	201	49.46	13.40	20.00	90.00	.27
N494	200	16.73	34.67	5.00	420.00	2.07
N503	201	15.69	6.83	3.00	60.00	.44
P549	201	14.23	6.61	1.0	60.00	.46
P554	199	18.48	17.55	5.00	260.00	.95
P555	200	39.62	22.64	4.00	300.00	.57

Table 11. Summary Statistics for the Final Edited Task Productive Capacity Measures

Task	n	Mean	S.D.	Minimum	Maximum	CV
E120	201	.24	.18	.01	.99	.76
E143	201	.28	.19	.01	.99	.67
F153	194	.45	.16	.01	.99	.36
F154	201	.38	.18	.01	.99	.46
F155	200	.32	.19	.01	.99	.59
F157	200	.34	.18	.01	.99	.51
F162	200	.35	.16	.01	.99	.46
G171	200	.26	.15	.01	.99	.58
G179	196	.29	.19	.01	.99	.67
G181	199	.26	.20	.01	.99	.76
H202	201	.50	.15	.01	.99	.31
H203	200	.56	.21	.01	.99	.37
H209	198	.38	.20	.01	.99	.53
H215	200	.45	.22	.01	.99	.49
H236	200	.30	.22	.01	.99	.72
H237	200	.35	.21	.01	.99	.59
H238	199	.35	.16	.01	.99	.45
I247	200	.24	.17	.01	.99	.72
I248	200	.40	.18	.01	.99	.45
I251	195	.25	.19	.01	.99	.75
I255	199	.26	.18	.01	.99	.72
I260	201	.45	.18	.01	.99	.40
I264	200	.19	.14	.01	.99	.72
I275	200	.38	.16	.01	.99	.42
I283	201	.40	.17	.01	.99	.42
I284	200	.19	.13	.01	.99	.71
I286	200	.37	.18	.01	.99	.49
I299	201	.13	.19	.01	.99	1.43
I300	201	.34	.19	.01	.99	.57
J332	198	.15	.15	.01	.99	1.02
J340	198	.21	.12	.01	.99	.57
J347	198	.31	.18	.01	.99	.60
J355	198	.40	.19	.01	.99	.48
L406	200	.28	.19	.01	.99	.70
L421	199	.35	.21	.01	.99	.60
L436	196	.39	.18	.01	.99	.46
L437	200	.32	.20	.01	.99	.62
M444	196	.12	.12	.01	.99	.98
M446	200	.47	.17	.01	.99	.35
M447	195	.36	.19	.01	.99	.53
N475	184	.35	.19	.01	.99	.55
N477	201	.37	.19	.01	.99	.53
N486	201	.20	.18	.01	.99	.91
N487	201	.44	.17	.01	.99	.39
N488	201	.36	.20	.01	.99	.54
N494	200	.42	.19	.01	.99	.45
N503	201	.28	.20	.01	.99	.74
P549	201	.29	.19	.01	.99	.65
P554	199	.48	.13	.01	.99	.27
P555	200	.26	.12	.01	.99	.46

Table 12. Average Percent Time Spent and Computed Task Weights by Duty Area

Duty Area	Skill Level			Task Weight
	3	5	7	
E	6%	10%	17%	11
F	18%	14%	7%	13
G	6%	5%	2%	4.33
H	13%	12%	7%	10.67
I	14%	13%	7%	11.33
J	5%	5%	2%	4
L	3%	3%	2%	2.67
M	3%	4%	2%	3
N	11%	8%	3%	7.33
P	8%	6%	4%	6

of summary statistics of the raw times indicates the coefficient of variation was greater than one for 13 tasks. It is thus apparent that the editing was effective in controlling the effects of outliers and getting the variance to more reasonable levels.

4.2 Selection of a Task Weighting Scheme.

After computing the task-level PCs, it was possible to weight them according to the weighting scheme described in Section 3.3.2. Recall that the weighting scheme actually applied weights to each job duty area, and all tasks from a particular duty area were assigned the same duty area weight. Further recall that the weights were based on the relative amount of time airmen spend doing particular types of tasks.

Table 12 shows the average percent time spent on each represented duty area broken out by each represented skill level. It also shows the computed weights by duty area. Again, the weights were an average of the average percent time spent across skill levels

4.3 Aggregation of the Task-Level Data into an Overall Productive Capacity Measure.

Once the task-level PC measures were computed and the task weights derived, it was possible to compute the aggregate or overall PC measure. Recall that aggregate PC was defined as a

Table 13. Aggregate Productive Capacity Measures Created

Variable	Description
PC_{uwavg}	Unweighted average of the final edited task-level PCs per individual.
PC_{wavg}	Weighted average of the final edited task-level PCs per individual.
$\widehat{PC}_{uwavg}$	Predicted value of the unweighted average, PC_{uwavg}.
$\widehat{PC}_{wavg}$	Predicted value of the weighted average, PC_{uwavg}.

Table 14. Summary Statistics for the Aggregate Productive Capacity Measures

Variable	n	Mean	S.D.	Minimum	Maximum
PC_{uwavg}	201	.33	.10	.02	.66
PC_{wavg}	201	.34	.10	.02	.68
$\widehat{PC}_{uwavg}$	169	.32	.04	.14	.36
$\widehat{PC}_{wavg}$	169	.32	.04	.15	.37

weighted average of the task-level PCs for an individual (see Figure 17). Also recall that a simple unweighted average was computed for comparative purposes. Table 13 provides a brief description of each aggregate variable created, for further reference.

The *predicted* values described in Table 13 were obtained from the estimated regression functions which are discussed in the next section.

Table 14 provides some summary statistics and Figure 21 provides histograms for the aggregate variables to give some insight into their distributions. Also, Table 15 shows the Pearson correlation coefficient between the weighted and unweighted versions of the variables.

Table 15. Correlation Between the Weighted and Unweighted Aggregate Productive Capacity Measures

Unweighted Variable	Weighted Variable	Correlation Coefficient
PC_{uwavg}	PC_{wavg}	$> .99^a$ ($n = 201$)
$\widehat{PC}_{uwavg}$	$\widehat{PC}_{wavg}$	$> .99^a$ ($n = 169$)

Superscript a indicates significance at the $\alpha = .05$ level.

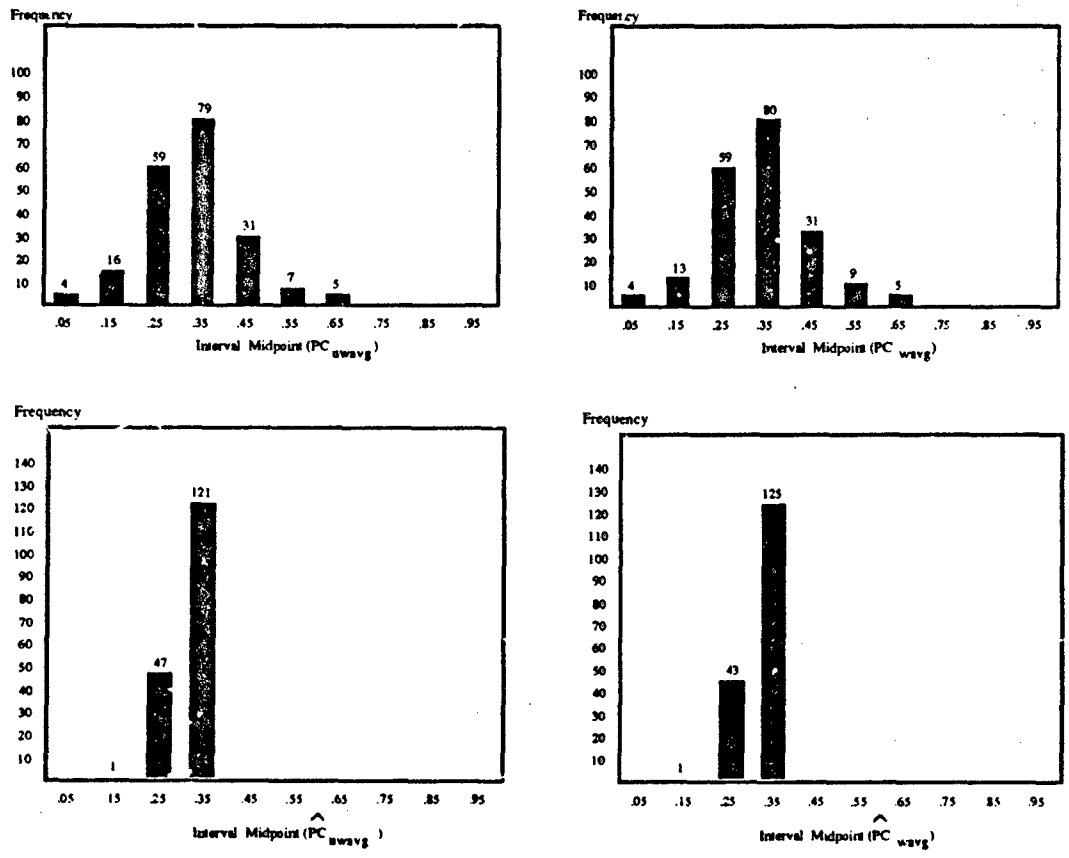


Figure 21. Histograms of the Aggregate Productive Capacity Measures

From the histograms of the aggregate variables (see Figure 21) it appears that the predicted values are negatively skewed. This is a reflection of both the shape of fitted response curves and the experience and aptitude levels of the sample. In other words, most of the sample possessed levels of the predictor variables which corresponded to the higher response points on the fitted response surface. Again, the reader is referred to the next section for discussion of the fitted response surfaces.

The summary statistics (see Table 14) and the Pearson correlation coefficients between the weighted and unweighted versions of the aggregate variables (see Table 15) indicated that the task weighting had negligible effects on the aggregate variables. The highly similar statistics and correlation coefficients very near one suggest that the weighted and unweighted versions of the variables are measuring approximately the same attributes. Thus, it appears that the weighting scheme and subsequent weighted averaging were ineffective in defining overall PC beyond what could be offered by simple averaging.

4.4 Development of Prediction Models.

As previously mentioned, full second-order logistic models were fit to PC both at the task level and at the aggregate level. Recall that the logistic model was linearized through formation of the logit response function, and the parameters were then estimated through ordinary least squares estimation. Table 16 summarizes the results of the logistic model regressions for the tasks.

Table 16: Regression Results for the Full Second-Order Logistic Models at the Task-Level

Task	$\hat{\beta}_0$ *Intercept*	$\hat{\beta}_1$ $(\times 10^{-4})$ Apt^2	$\hat{\beta}_2$ $(\times 10^{-4})$ Exp^2	$\hat{\beta}_3$ $(\times 10^{-4})$ $Apt \times Exp$	$\hat{\beta}_4$ *Apt*	$\hat{\beta}_5$ *Exp*	F_0	R^2
E120	-1.70	4.49	-2.02^a	-5.11^a	-.040	$.064^a$	4.83^a	.13
E143	-3.57	-1.57	-1.96^a	-1.92	.037	$.038^a$	2.44^a	.07
F153	-.489	.636	-1.39^a	1.47	-.010	.012	2.86^a	.08
F154	-8.22^a	-10.0^a	-1.15^b	-3.83^b	$.169^a$	$.053^a$	5.04^a	.13
F155	-6.07^b	-5.82	-2.50^a	-1.89	.101	$.051^a$	3.96^a	.11
F157	-3.42	-3.88	-.633	-1.13	.059	.021	.98	.03
F162	-2.56	-1.91	-1.68^a	.110	.031	$.025^b$	2.98^a	.08

Table 16: (continued)

Task	β_0 *Intercept*	β_1 ($\times 10^{-4}$) *Apt*^{2}	β_2 ($\times 10^{-4}$) *Exp*^{2}	β_3 ($\times 10^{-4}$) *Apt* × *Exp*	β_4 *Apt*	β_5 *Exp*	F_0	R^2
G171	-4.61	-3.34	-.881	-3.13	.066	.038^b	1.06	.03
G179	-4.19	-3.48	−3.91^a	3.00	.048	.035	4.86^a	.13
G181	2.73	10.3	-1.76	-.002	-.144	.029	1.87	.05
H202	−5.04^b	-6.38	-.100	-1.98	.173^b	.022	1.79	.05
H203	-5.46	-5.99	−1.53^b	2.31	.116	.040^b	2.31^a	.07
H209	-1.96	-1.80	-1.42	.646	.027	.012	.63	.02
H215	-1.59	1.04	−2.27^a	-1.12	.001	.040	1.65	.05
H236	−9.23^a	−11.7^b	−2.24^a	-.838	.189^b	.037	2.76^a	.08
H237	−6.40^b	-4.46	-.401	−5.77^a	.108	.048^a	2.15^b	.06
H238	-2.92	-2.58	1.91^a	-1.19	.053	-.007	3.02^a	.09
I247	-2.97	-2.48	-.940	2.64	.037	-.007	.93	.03
I248	-4.03	-4.23	−1.69^a	-.138	.072	.025	2.65^a	.08
I251	-1.06	1.07	−2.90^a	2.93	-.024	.013	2.22^b	.07
I255	-4.67	-3.45	−2.93^a	.757	.060	.031	3.54^a	.10
I260	1.22	3.25	−2.04^a	2.94	-.051	.001	3.50^a	.10
I264	-3.99	-2.79	−2.41^a	-.037	.042	.034^b	2.70^a	.08
I275	-1.75	-1.63	−1.49^a	.867	.021	.016	2.17^b	.06
I283	.083	-.022	−1.49^b	1.88	-.013	.005	.86	.03
I284	-2.91	-2.09	−1.61^a	3.32	.023	.002	2.42^a	.07
I286	−5.81^b	-6.74	−1.29^b	-1.02	.113	.031	2.43^a	.07
I299	−10.68^a	−14.0^b	−3.29^a	2.01	.207^b	.026	2.76^a	.08
I300	-3.46	-3.81	-.733	2.50	.059	-.004	1.77	.05
J332	−8.22^a	-9.22	−2.47^a	.428	.144	.031	3.28^a	.09
J340	-1.59	2.79	−1.10^a	-2.12	-.025	.029^a	2.45^a	.07
J347	-2.25	2.47	−2.45^a	-.140	-.015	.037	3.42^a	.10
J355	-2.08	.160	−1.59^a	-1.26	.012	.032^b	1.88	.06
L406	-5.34	-5.55	−1.97^a	.089	.091	.024	1.98^b	.06
L421	-3.28	-2.88	−2.61^a	- 473	.046	.037^b	3.00^a	.09
L436	3.33	6.31	-.552	1.01	-.100	-.005	.500	.02
L437	−11.49^a	−16.1^a	-.944	-4.36	.258^a	.052^a	3.17^a	.09
M444	-.273	3.83	−1.74^a	1.94	-.063	.007	1.08	.03
M446	-2.57	-2.88	−2.22^a	1.73	.042	.023	5.04^a	.13
M447	-5.11	-5.65	-.932	-1.07	.099	.020	1.05	.03
N475	−6.93^a	−10.5^b	-1.17	.970	.156^b	.012	1.59	.05
N477	2.76	6.45	−2.10^a	2.79	-.107	.012	2.87^a	.08
N436	-6.14	-3.07	-.130	−10.3^a	.081	083^a	2.44^a	.07
N487	-2.86	-1.52	-.817	-1.71	.043	023	1.78	.05
N488	2.49	7.19	−1.65^a	-.295	-.103	.020	1.36	.04
N494	-3.22	-2.40	-.401	-1.78	.051	.027	1.77	.05
N503	-3.95	-1.40	-.818	-3.57	.040	.045^a	1.97^b	.06
P549	-4.15	-5.02	-1.04	-.181	.078	.011	.49	.01
P554	-2.82	-4.23	-.371	.067	.066	.008	1.03	.03
P555	.632	4.92	−1.39^a	-.481	-.068	.022	2.09^b	.06

Superscript b indicates significance at the $\alpha = .10$ level

Superscript a indicates significance at the $\alpha = .05$ level

The regression results in Table 16 for the tasks indicate some consistent results. First, the R^2s were consistently low, ranging from .01 to .13. Second, aptitude did not appear to have much influence on task PCs as indicated by the statistical significance of the corresponding parameters. The associated aptitude coefficients, β_1, β_4, or both, tested significantly different from zero at the $\alpha = .05$ level for only two tasks. Third, experience seemed to be more strongly related to PC with either β_2, β_5, or both, testing significantly different from zero at the $\alpha = .05$ level for 33 of 50 tasks. It is important to note that the aptitude/experience interaction coefficient, β_3, tested significantly different from zero for only four tasks. Overall, these results were in partial agreement with those of Schmidt and others and Alley and others (1) (36) . They also found that there does not appear to be an aptitude/experience interaction affecting job performance. However, they found aptitude to be an important determinant of job performance.

Overall, the task-level logistic models for predicting PC did not perform as well as AL/HRM's task-level learning curve models for predicting untransformed estimated times (38). More of the learning curve models (41 of 50) were significant and they yielded generally higher R^2s (ranging from .01 to .20). But, as mentioned, learning curve models are only useful for determining how fast a piece of work can be completed given the worker's aptitude and experience level. A transformation must still be applied to the time data to provide a standardized, interpretable work output measure like PC. A second drawback of the learning curve model is that it is difficult to develop a meaningful model for predicting overall performance measures when the appropriate level of job specificity for data collection is the task level. There seems to be no meaningful way of aggregating task-level performance times into an overall measure that could be predicted.

Because of the large number of tasks studied, detailed residual analyses to check model aptness were not performed at the task level. An aptness analysis was performed for the model for predicting the aggregate measure.

Table 17. Regression Results for the Full Second-Order Logistic Model at the Aggregate Level

Variable	β_0 *Intercept*	β_1 ($\times 10^{-4}$) Apt^2	β_2 ($\times 10^{-4}$) Exp^2	β_3 ($\times 10^{-4}$) $Apt \times Exp$	β_4 *Apt*	β_5 *Exp*	F_0	R^2
PC_{wavg}	-2.97^a	-2.10	-1.24^a	-.610	.040	$.023^a$	6.07^a	.16

Superscript a indicates significance at the $\alpha = .05$ level.

Finally, in reference to the task-level models, recall that they were run primarily as a screening exercise to provide insight as to whether any model terms from the second-order model could be dropped. The task-level models obviously indicated that the terms including the aptitude variable were potential candidates for removal from the model. The aggregate model was analyzed, in part, to further explore this possibility.

Table 17 provides the regression results for the aggregate variable, PC_{wavg}, regressed on aptitude and experience using the linearized logit response function. Table 17 contains some interesting results. In predicting the aggregate measure, experience seemed to be an influencing factor. This was indicated by β_2 and β_5 both testing significantly different from zero. The aptitude coefficients, β_1 and β_4, tested not significantly different from zero. Overall, the results of the aggregate model paralleled the results of the task models in that experience was an influencing factor, but aptitude and the aptitude/experience interaction were not. These results again are in partial agreement with those of Schmidt and others and Alley and others (1) (36). They found no interaction effects but in contrast, they did find significant aptitude effects. In comparison to the Air Force's other PC studies, the aggregate model R^2 was comparable to those found for the AGE specialty by Faneuff and others (13:10). They reported R^2s of .17 and .20 using the ASVAB E and M scores as aptitude variables, respectively. But, the R^2s of the current study were much lower than that for the aggregate model of Carpenter and others (R^2=.44) for specialty 328X0 (5:22).

As mentioned, the results of the regression using the full second-order logit response model for PC_{wavg} showed that none of the parameters for terms which included the aptitude measure tested significantly different from zero. This was further indication that the aptitude predictor was

Table 18. Forward Stepwise Regression Results for the Second-Order Logistic Model at the Aggregate Level

Variable	β_0	β_1 $(\times 10^{-4})$	β_2 $(\times 10^{-4})$	β_3 $(\times 10^{-4})$	β_4	β_5	F_0	R^2
	Intercept	Apt^2	Exp^2	$Apt \times Exp$	*Apt*	*Exp*		
PC_{wavg}	-1.23^a	0	-1.31^a	0	0	$.019^a$	12.69^a	.13

Superscript a indicates significance at the $\alpha = .05$ level.

Table 19. ANOVA Table for the Aggregate Productive Capacity Measure after Forward Stepwise Regression

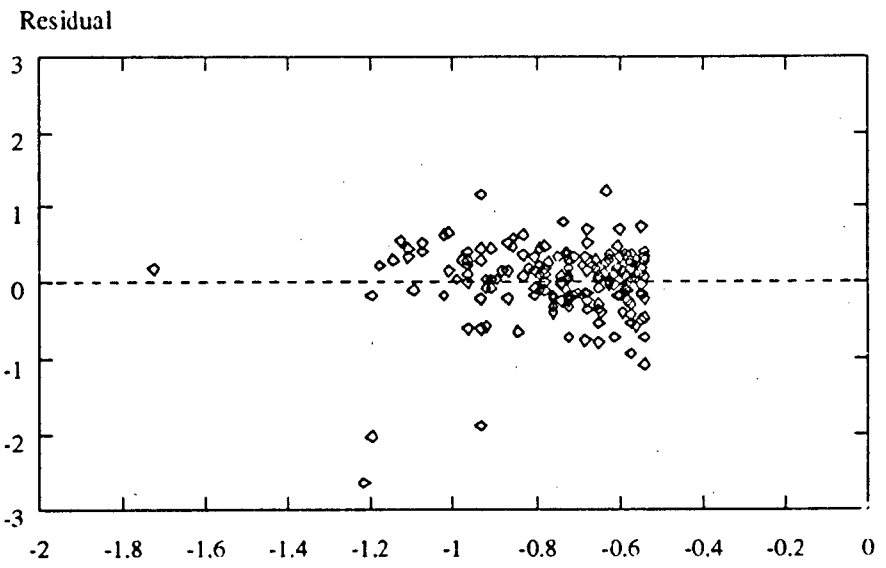
Source of Variation	SS	df	MS	F_0
Regression	5.73	2	2.87	12.69^a
Error	37.50	166	.23	
Total	43.23	168		

Superscript a indicates significance at the $\alpha = .05$ level

a candidate for removal from the model. A forward stepwise regression was then run beginning with the full second-order model to determine if the aptitude terms could be dropped. The criterion for a term's entry into the model was F statistic significance at the $\alpha = .05$ level. The same criterion was used for a term's departure from the model. The forward stepwise regression did in fact drop all terms involving the aptitude variable from the model. Table 18 provides the results of the stepwise regression and Table 19 provides the final ANOVA table .

The model after stepwise regression was selected as the final model, provided that it would meet the linear model assumption of normality of error terms ($\epsilon \sim N(0, \sigma^2)$). Figure 22 provides the results of an aptness analysis for the final model to check the normality assumption. The figure includes a plot of the model residuals vs. fitted values a normal probability plot of the residuals.

The top plot in Figure 22, a plot of the residuals vs. the fitted values, shows a fairly even band of points around the zero-residual line. This indicated that the variance of the residuals and thus the variance of the actual error terms is fairly constant across differing levels of the predicted values. This homoscedasticity is, of course, desirable. If the error variance was not constant across



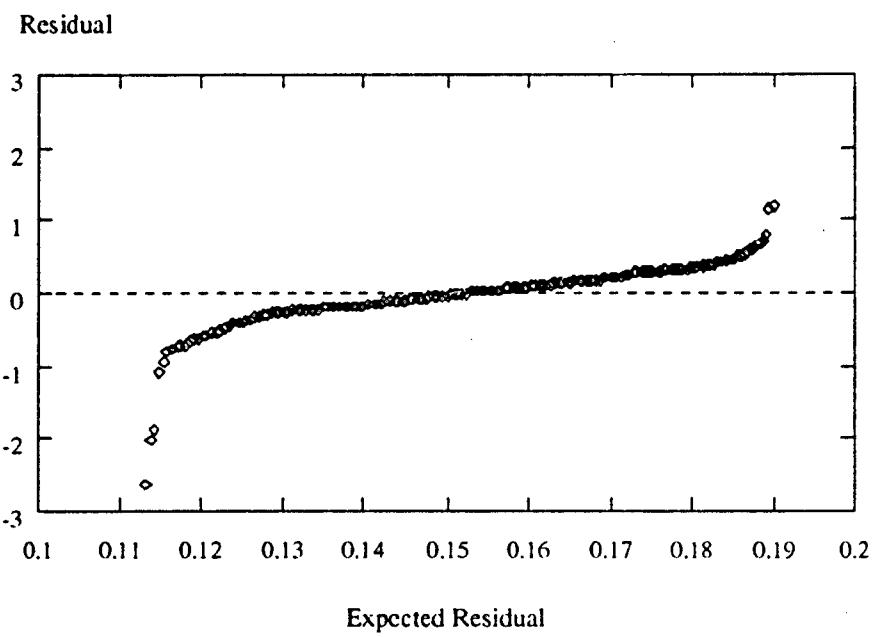


Figure 22. Residual Plots as an Aptness Analysis of the Fitted Model

Table 20. Correlation Between the Aggregate Productive Capacity Measures and Other Job Performance Measures

	JKT	GPC	MTPC
PC_{wavg}	.08 (n = 193)	.44[a] (n = 199)	-.13 (n = 58)
$\widehat{PC}_{wavg}$	.22[a] (n = 169)	.25[a] (n = 167)	−.27[b] (n = 51)

Superscript a indicates significance at the α = .05 level.
Superscript b indicates significance at the α = .10 level.

the different levels of the fitted values, then the model would not be appropriate for the fitted region since it is assumed that the error terms, ϵ, are distributed $N(0,\sigma^2)$. The bottom plot, a normal probability plot, shows a high degree of linearity toward the center of the data with a few outliers at each end showing clear nonlinearity. Linearity is desirable because it indicates the actual residuals and their expected values under the normal assumption are highly correlated. Linearity implies normality of the residuals and thus normality of the model error terms. The nonlinearity at the ends of the plot was not overly worrisome since it is due to a relatively small number of outlying points. Overall, the conclusion was that the fitted second-order logistic model with aptitude excluded was appropriate for the data.

4.4.1 Correlational Analysis of the Estimated Model Results. After the aptness analysis, predicted values of PC_{wavg} ($\widehat{PC}_{wavg}$) were obtained from the final fitted model. As part of the model assessment, it was determined that the correlation between the final model's predicted values and other job performance measures would offer insight as to the model's effectiveness. Table 20 shows the correlation between the computed and predicted aggregate variables and other previously defined job performance measures collected under the Productive Capacity Project.

Table 20 indicates that the aggregate variable computed from the task-level PC measures, PC_{wavg}, correlated more highly with GPC than did the associated predicted variable, $\widehat{PC}_{wavg}$. This is not terribly surprising since GPC, like PC_{wavg}, is also the result of supervisor estimation.

Table 21. Correlation Matrix of the Other Job Performance Measures

	JKT	GPC	MTPC
JKT	1.0[a] (n = 196)		
GPC	.12 (n = 191)	1.0[a] (n = 199)	
MTPC	−.44[a] (n = 60)	-.18 (n = 57)	1.0[a] (n = 60)

Superscript a indicates significance at the $\alpha = .05$ level.

The predicted values, $\widehat{PC}_{wavg}$, correlated more strongly with the objectively-derived measures, JKT and MTPC.

There seemed to be a pattern of higher correlation between the predicted values and the more objective measures. A similar pattern existed between the computed average measure, PC_{wavg}, and the more subjective measure, GPC. This seemed to indicate that the subjectively-derived measures are measuring different dimensions of performance than the predicted variable and the objective variables.

One final noteworthy finding is the relatively low correlation between mean PC derived from actual stopwatch times (MTPC) and computed average PC derived from supervisor estimates (PC_{wavg}). This is an indication that the supervisors' ratings may be measuring a different dimension of performance than the actual stopwatch times, or a great deal of *noise* resulting from rating biases of the supervisors.

To summarize the results of the correlational analysis, PC_{wavg} correlated more strongly with GPC than did the associated predicted values. This seemed to indicate that predicted values, and thus the model, captured less of global PC (as judged by the supervisors in their GPC ratings) than the computed average data. This may be an indication that the model is not measuring what it is supposed to—overall PC. But, on the other hand, the predicted values did correlate more highly with the other objective measures indicating that the model is predicting job performance in at

least one respect. The assessment of the model through correlational analysis thus gives conflicting results.

Table 21 was included simply to give the reader an indication of how the other job performance measures relate to one another.

4.4.2 Graphical Representation of the Estimated Logistic Models. The preceding regression results and correlational analyses were helpful in providing insight as to how the predictors potentially influence PC, and how the aggregate variables (computed and predicted) relate to other job performance measures. This section is intended to provide additional insight into the estimated model by providing a graphical representation of the fitted models. Figure 23 shows the fitted response curve for the final model to provide a graphical representation of the relationship between experience and PC. It is plotted over the effective range of the predictor, experience (one to 170 months). Figure 24 shows the plotted surface for the full second-order model prior to the stepwise regression, to show the relatively mild effects of aptitude and interaction on estimated PC. Recall that in the stepwise regression, the aptitude terms were dropped. The full model is likewise plotted over the effective range of predictors, aptitude (M score 45–99) and experience (one to 170 months).

The fitted response surfaces were obtained by entering the logistic model parameter estimates into the logistic model function. Equation 27 shows the equation for the final model, and Equation 28 shows it for the full second-order model.

$$\widehat{PC}_{\text{Weighted Final Model}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_2 x_2^2+\hat{\beta}_3 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_2 x_2^2+\hat{\beta}_3 x_2}} \tag{27}$$

where

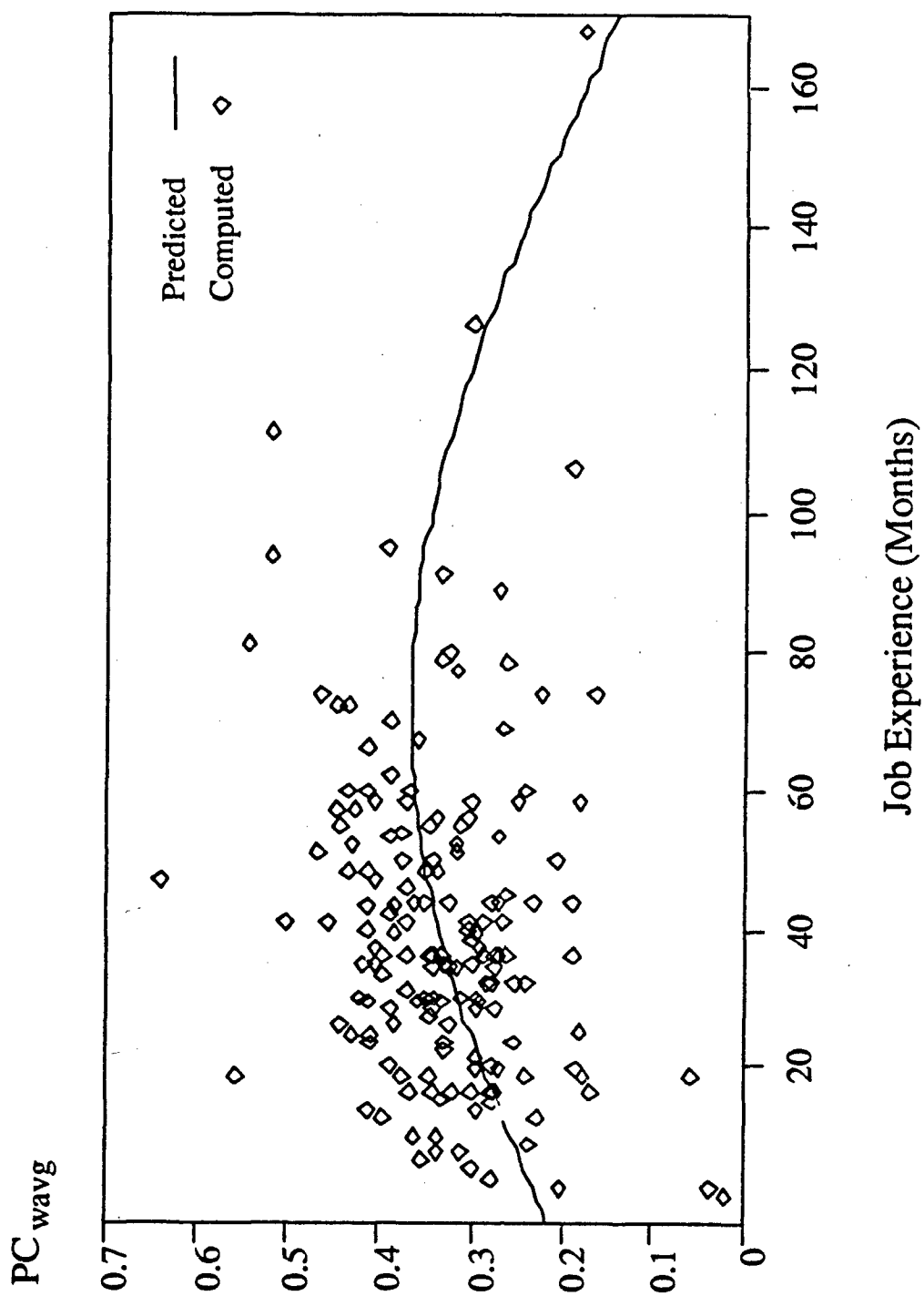


Figure 23. Fitted Response Curve and Scatterplot for Productive Capacity Over the Effective Range of Experience

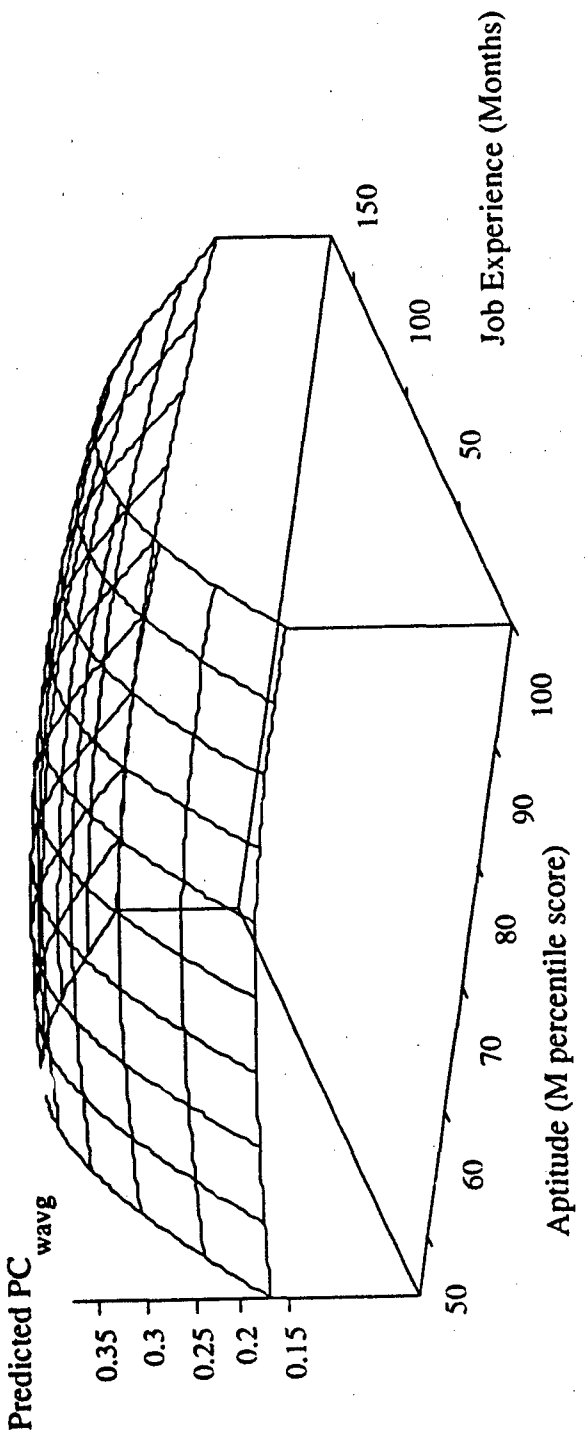


Figure 24. Fitted Response Surface for Productive Capacity Over the Effective Range of Aptitude and Experience

$\widehat{PC}_{wavg_{Final\ Model}}$ = *predicted weighted average productive capacity from the final model after the stepwise procedure*

x_2 = *months of job experience*

$\hat{\beta}_0$ = -1.234482

$\hat{\beta}_2$ = $-.000131$

$\hat{\beta}_5$ = $.019038$.

$$\widehat{PC}_{wavg_{Full\ Model}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1^2+\hat{\beta}_2 x_2^2+\hat{\beta}_3 x_1 x_2+\hat{\beta}_4 x_1+\hat{\beta}_5 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1^2+\hat{\beta}_2 x_2^2+\hat{\beta}_3 x_1 x_2+\hat{\beta}_4 x_1+\hat{\beta}_5 x_2}} \tag{28}$$

where

$\widehat{PC}_{wavg_{Full\ Model}}$ = *predicted weighted average productive capacity from the full second-order model*

x_1 = *ASVAB Mechanical percentile score*

x_2 = *months of job experience*

$\hat{\beta}_0$ = -2.969180

$\hat{\beta}_1$ = $-.000210$

$\hat{\beta}_2$ = $-.000124$

$\hat{\beta}_3$ = $-.000061$

$\hat{\beta}_4$ = $.039573$

$\hat{\beta}_5$ = $.022894$.

To increase the interpretability of the fitted response curve and surface, the entire surfaces were rescaled to zero/one space much like the edited task-level PC values. This was to ensure a minimum predicted PC of zero and a maximum of one so that PC could be interpreted as a proportion of maximum possible output. Equation 29 mathematically shows the equation for the

rescaled surfaces. The rescaled response surfaces are shown in Figure 25 and Figure 26 for the final and full models, respectively.

$$\widehat{PC}_{wavg_{Rescaled}} = \frac{\widehat{PC}_{wavg} - \widehat{PC}_{wavg_{min}}}{\widehat{PC}_{wavg_{max}} - \widehat{PC}_{wavg_{min}}} \tag{29}$$

where

$\widehat{PC}_{wavg_{Rescaled}}$ = *rescaled predicted average mean productive capacity*

$\widehat{PC}_{wavg_{max}}$ = *Minimum value of* $\widehat{PC}_{wavg}$

$\widehat{PC}_{wavg_{min}}$ = *Maximum value of* $\widehat{PC}_{wavg}$.

The plotted response curves and surfaces all show PC initially increasing with experience until it reaches a maximum, and then begins to steadily decrease. In the case of the plotted surface for the full second order model, this is shown to occur at all levels of aptitude. The plots for the full model also show PC generally increasing with aptitude at all levels of experience. There does appear to be a peak and a slight decrease in PC with increasing aptitude. Once again, in reference to the plot of the full model, very little interaction was present, as indicated by the fairly constant effects of one predictor with varying levels of the other.

Before drawing conclusions, it is important to recall that the models did not fit the data very well (for the full model R^2 = .16, and for the final model R^2 = .13). Also, recall from Table 9, the two-way distribution of aptitude and experience, that there were relatively few data points indicating experience beyond 96 months. The model must thus be interpreted cautiously beyond this point. These two facts suggest that the response curves and surfaces should not be viewed with exactness, but in general terms. They should serve only to provide some possible insight as to how the factors might effect eachother.

The decreasing PC with increasing experience over a portion of the curves and surfaces was an unexpected result. This seemed to indicate that there is some point in an airman's career

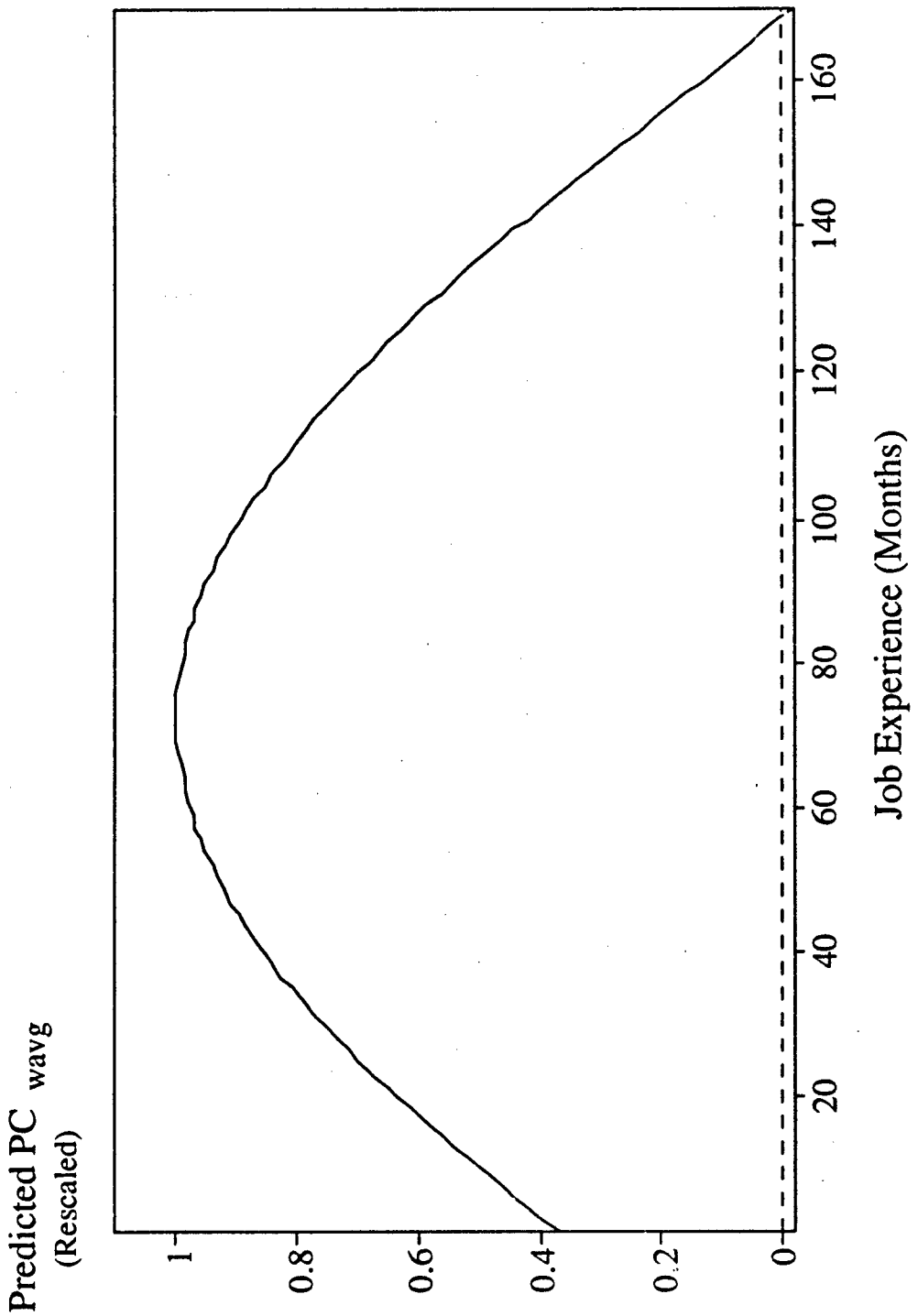


Figure 25. Rescaled Fitted Response Curve for Productive Capacity Over the Effective Range of Experience

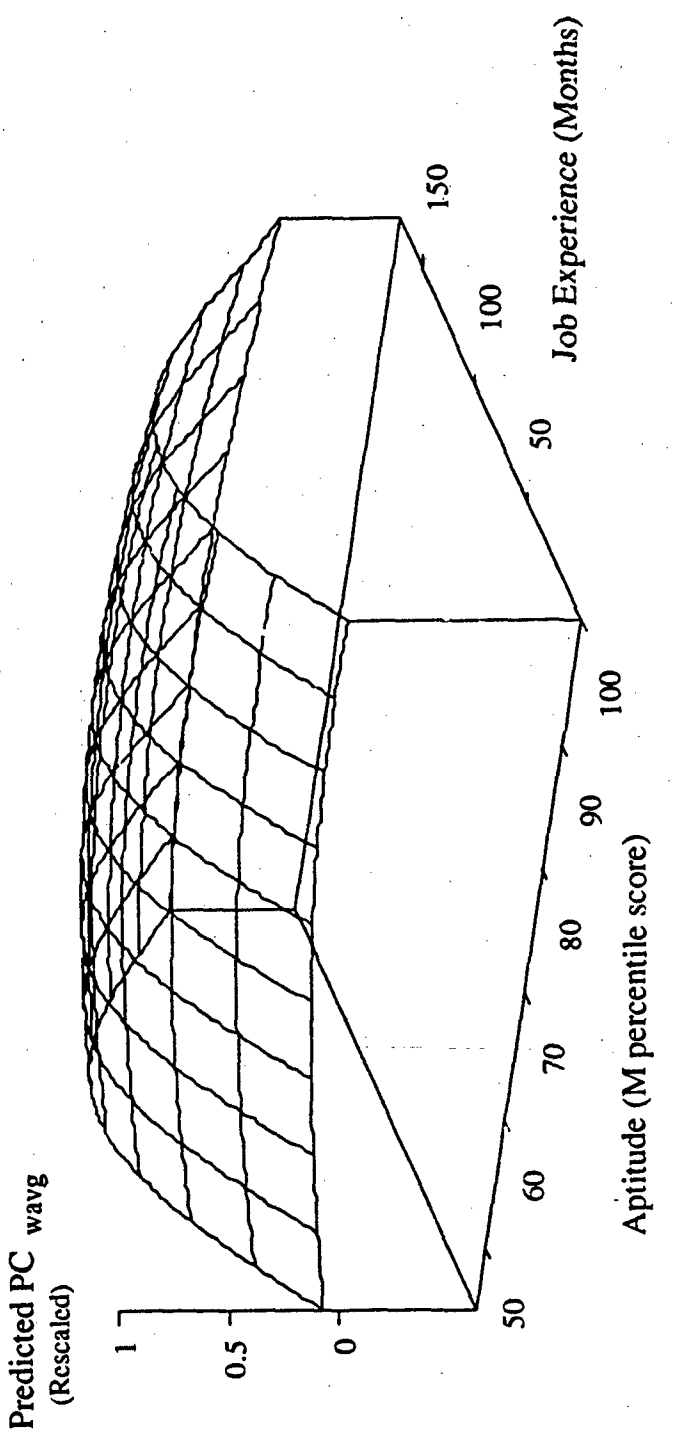


Figure 26. Rescaled Fitted Response Surface for Productive Capacity Over the Effective Range of Aptitude and Experience

where he or she may begin to experience skill degradation, or decreasing PC on the types of tasks studied. The estimated logistic function for the final model was put into GINO (*G*eneral *IN*teractive *O*ptimizer) to identify exactly where the maximum PC point on the curve occurs, and thus where the performance degradation may begin (22). Maximum PC occurs on the surface at about 71 months of job experience. This seemed to indicate that after approximately six years of job experience, the capacity to perform hands-on production-type job tasks decreases for 454X1 personnel. This result might be explained by the fact that Air Force enlisted personnel typically begin to make the transition into supervisory roles at around the six year point. This means they begin to spend less time practicing production-type tasks so skill degradation might reasonably be expected. This is not to say that an airman's overall performance decreases after the six year point, only performance on the types of tasks studied under the Productive Capacity Project. Hands-on performance on such types of tasks becomes decreasingly important as airmen advance in grade and move on to supervisory roles. A more appropriate measure of performance for more senior members would most likely have to include measures of their ability to supervisor.

A final point to be made concerning Figure 24 and Figure 26 for the full model is that maximum PC occurred near an aptitude score of 84, according to GINO. In looking at the plotted surfaces, there does not seem to be a significant decrease in performance beyond this score. There is simply no strong indication that PC truly does peak and then decrease with increasing aptitude. Since the model provided significantly less than a perfect fit, it may simply be enough to note that PC tends to increase with aptitude, in general, at all levels of experience.

V. Summary, Conclusions and Recommendations

Recognizing that the Air Force could greatly benefit from acquiring the ability to forecast the future job performance of its personnel, this research effort set out to develop experimental, descriptive regression models for predicting the job performance of personnel in specialty 454X1, Aerospace Ground Equipment. Hopefully, this modeling activity will serve to help Air Force planners take another step in their iterative and on-going quest for adequate job performance models.

The research objectives, as presented in Chapter 1, were as follows:

1. Formulate a Productive Capacity Measure from Estimated Task Performance Times
2. Select a Task Weighting Scheme
3. Aggregate the Task-Level Data into an Overall Productive Capacity Measure
4. Develop Prediction Models

Data for the analyses were collected by the Air Force under its Productive Capacity Project (21). The primary dependent (response) variables were raw estimated task performance times for airmen in specialty 454X1, and the independent (predictor) variables were mechanical aptitude and job experience.

The following sections provide a brief recapitulation of the research methods used to meet these research objectives and contain a summary of conclusions and recommendations for further research.

5.1 Summary and Conclusions.

5.1.1 Formulating a Productive Capacity Measure from Estimated Task Performance Times.

The primary response data analyzed were raw estimated task performance times for 204 airmen in specialty 454X1, Aerospace Ground Equipment. The estimated times were provided by the airmens' supervisors for 50 job tasks commonly performed by personnel in the specialty. An initial

research objective was to determine how to transform the task-level time data into measures that are interpretable and able to be aggregated across tasks. At the task level, an interpretable measure, PC, was formulated according to the method proposed by Carpenter and others, t^*/t (5:21). In the formulation, t^* represented an estimate of the fastest possible time in which a given task could be completed, and t represented the estimated time for an airman to complete that task. The measure can be interpreted as an individual's output as a proportion of maximum possible output.

Several considerations had to be accounted for in computing task-level PC. Most importantly, the raw estimated performance times from which the PC measures were derived tended to be highly variable with an often unbelievable range of values within a task. This indicated a need for editing to control for influential outliers. As a result, several stages of data editing were applied to the raw estimated times and to the computed PCs to obtain reasonable distributions of the task-level PCs.

5.1.2 Selecting a Task Weighting Scheme. Since PC is a quantity-based measure of work output capability, it seemed appropriate to weight the tasks by the relative amount of time airmen spend doing them on average. This was to account for the fact that airmen may spend varying amounts of time on different tasks, some of which they are productive on, and some on which they are not. Tasks were weighted by a factor derived through averaging *Average Percent Time Spent Performing Duties* data (collected by the Occupational Measurement Squadron) across relevant skill levels. Duty area weights were applied to tasks from that area. Greater weight went to those tasks performed most frequently.

The applied weighting scheme had little effect on the computed aggregate variables. The weighted average measures, when compared to their unweighted counterparts, had highly similar descriptive statistics. The weighted and unweighted versions of the variables were also highly correlated. The conclusion is that the applied weighting scheme had no noticeable effect on aggregate PC.

5.1.3 Aggregating the Task-Level Data into an Overall Productive Capacity Measure. At the overall or aggregate level, PC was defined and computed as a weighted average of the task-level PC values for each experimental subject. Along with weighted averaging, the task-level PC data were aggregated through simple, unweighted averaging for comparison purposes. The need for aggregation existed because overall measures are of more importance in the bigger scheme of manpower modeling and planning. Task-level information is important but manpower decisions usually cannot be made based on an individual's predicted performance on single tasks. Also, modeling at the task level for each of the approximately 250 AFSs would simply be too cumbersome. Because jobs tend to be multifaceted and dynamic, and because task-level modeling is potentially too burdensome, it was desirable to compute and model an aggregate measure.

5.1.4 Developing Prediction Models. Both task-level and aggregate PC measures were regressed on aptitude and job experience using a second-order logistic model. The aptitude variable used was the Mechanical percentile score from the ASVAB obtained by each subject upon applying for enlistment. The experience variable used was the subjects' self-reported job experience at the time the estimated times were collected.

At the task-level, R^2s were consistently low for the logistic model, ranging from .01 to .13. This may indicate that there are other predictor variables influencing PC that were not addressed in this thesis. Another possible explanation of the low R^2s is that the assumption of validity and reliability of the PC data collection instrument and method is not sound. Supervisors are known to be subject to many types of biases which affect their judgements concerning the performance of their personnel (6:82–84). The low R^2s may be indicative of the fact the supervisors are introducing *noise* into the data from such biases, and thus adversely affecting validity and reliability, and thus model fit.

Residual analysis of the aggregate logistic model indicated that it was reasonably appropriate for the data. The model for predicting the aggregate measure yielded results that were comparable

to the task-level model results. Experience seemed to be a significant predictor while aptitude and the aptitude/experience interaction did not. The model R^2 for the full second-order aggregate logistic model was .16. The full second-order model was subjected to forward stepwise regression which indicated that all terms involving the aptitude variable could be dropped from the model. The final model involved a constant intercept term and linear and quadratic experience terms. The final model yielded an R^2 of .13.

After the logistic model parameters were estimated, predicted PC values were computed for the aggregate measure. These were correlated with other subjective and objective job performance measures collected under the Productive Capacity Project. The predicted values showed correlations significantly different from zero for each measure.

Fitted response surfaces for the estimated aggregate models were plotted and they indicated a pronounced peak for PC with respect to experience. There was some evidence that PC may begin to decrease for AGE personnel after about the six year point in their career. This may be reflective of skill degradation which may occur as airmen lose practice on hands-on type work as the transition to supervisory roles is made. It may also be the result of having only a few data points for higher levels of experience, or it may be simply an artifact of the relatively low degree of model fit.

Overall, the level of model fit (R^2) tended to be low, but comparable to that found for similar studies (5) (13) (38). R^2s of the current magnitude indicate that more work must be done to create more robust prediction models.

5.2 Recommendations.

The previous section provided a brief summary of the research objectives, methodology and findings. It did not, however, discuss the additional research questions which arose during the effort. As mentioned in the first chapter, exploratory or descriptive research such as this often spawns as

many research questions as it answers. This section will address some of the issues which came to the forefront in the current effort. These issues will be discussed in the context of recommendations for further research.

5.2.1 Formulating the Productive Capacity Measure. The current emphasis in Air Force job performance measurement is on using performance time data in deriving job performance measures. This is because most Air Force manpower modeling and planning involves performance criteria such as *sortie generation rates* and *mean time to repair aircraft*. Such measures are quantity-based and therefore indicate the need to assess and predict work output referenced to time. As a result, the Air Force is researching cost-effective methods for obtaining work performance time data.

As indicated in the current analyses, the current method of obtaining the performance time estimates (through free response supervisor estimation) yielded ranges of values which were excessively wide (see Table 10). This may be due in part to that fact that supervisors provided their estimates in a virtual free response format. This means that they were unconstrained in reference to the estimates they could make. In future studies, it is recommended that supervisors be forced to limit their time estimates to a pre-established reasonable range. A reasonable range of estimates could be derived using SMEs, much like Leighton and others used SMEs to develop benchmark times (21).

Other recommendations involve the formulation of PC measures from the task performance time data. One potential problem with creating PC measures according to the Carpenter and others formulation, t^*/t, is that the computed task-level PCs for each individual are in part based on the single task-level measure t^*. Since the computed PCs are based on them, care must be taken to obtain t^* values that are accurate so that the resulting PC values are properly interpretable. In the current research, it was pointed out that the raw estimated times, t, for each task tended to be highly variable indicating inconsistencies in the supervisors' opinions about what a reasonable

range of performance should be. This places some doubt in the accuracy of the estimated times, especially those near the fast and slow end of the estimated range. Since t^* was computed as the .99× (*minimum estimated time for the task (after editing)*), there is some question as to the credibility of such t^* values. An appropriate way to address this problem may be to compute the PC measure as in time studies where PC is computed as $(t_{avg}/t) \times 100$. Computed in this fashion, PC for a task is not dependent on a single measure t^*, but on the task average, t_{avg}.

Also, it is important to note that Carpenter and others' PC formulation is not a linear transformation of the time variable, t, from which it is computed. This is what prompted AL/HRM to formulate PC as t/t^*, an inversion of the Carpenter and others formulation (7). A nonlinear transformation can have the effect of influencing the degree of linear relationship between a variable and another. It is recommended that the nonlinearity introduced by the Carpenter and others' formulation be studied to determine its effect, and whether a linear transformation should be considered in future studies.

5.2.2 Selecting a Task Weighting Scheme. Because of the nature of the PC measure (quantity-based), it is recommended that *relative time spent* measures continue to be considered as a weighting factor. The PC measure, as defined, is indicative of a worker's output relative to some standard. In the current effort, that standard was t^*, an estimate of the fastest possible performance time. As such, the PC measure at the task-level must somehow be given different weights reflective of how often the tasks are performed. This is so that an aggregate measure which represents an airman's actual capacity to produce (given the average job scenario) can be computed. Recall that in the current effort, weights were derived for job duty areas as opposed to individual tasks. This was due to the unavailability of task-level data. It is recommended that an attempt be made to obtain and use *relative time spent* data derived for individual tasks as opposed to those for an entire duty area. This will further differentiate tasks on level of importance and may yield a more meaningful aggregate PC measure.

5.2.3 Aggregating the Task-Level Data into an Overall Productive Capacity Measure.

One problem with averaging (both weighted and unweighted) task-level measures is that there is significant information loss. In this thesis for instance, the actual response data for each individual was a row vector of about 50 task-level PC measures (see Figure 17). By weighted averaging, these were collapsed into a single measure. In collapsing the data, any unique information provided in individual task ratings was lost or dampened. Perhaps a reduction in the dimensionality of the response from 50 measures per person to one measure per person was too drastic.

One alternative to averaging is to treat the 204 × 50 (subjects × tasks) response matrix as a multivariate analysis problem. A common dimensionality-reduction technique that could be applied is factor analysis. According to Dillon and Goldstein,

> Factor analysis attempts to simplify complex and diverse relationships that exist among a set of observed variables by uncovering common dimensions or factors that link together seemingly unrelated variables, and consequently provides insight into the underlying structure of the data. (11.53)

In other words, factor analysis could be used to reduce the original set of 50 response variables to a smaller subset of factors that account for most of the variance in the task-level data (11:23). In factor analysis, a *factor* represents an underlying qualitative dimension like a coordinate axis, which defines the way in which different variables differ on that dimension (11:60). Factor analysis results in factor scoring coefficients which can be used to compute factor scores given known levels of the analyzed variables. Factor analysis basically takes advantage of the underlying correlational structure in the variables under analysis. Factors are derived such that correlated variables tend to *load* on the same factors. Factors, then, represent common dimensions that correlated variables share. For a more complete discussion of factor analysis, refer to Dillon and Goldstein (11).

The response matrix in the current study could be factor-analyzed to determine any factor structure that could be used to reduce the number of response variables to a set of less than 50 factors. Prediction models could then theoretically be developed to predict computed factor scores. Factor analysis seems to be a reasonable midpoint between collapsing the data into a single measure

through averaging, and modeling with task-level data. The analyst or manpower modeler would of course be left with the non-trivial task of interpreting the factors and resulting factor scores.

Another alternative for reducing the response matrix to less than 50 variables would be to compute aggregate measures at the duty area level. Referring to Table 4, there are 20 duty areas for the AGE specialty, 10 of which were represented by tasks in the current effort. The reduction from 50 task-level variables to 10 or 20 duty area variables would be substantial. Aggregating tasks from the same duty area, perhaps through weighted averaging, would provide aggregate variables representing reasonable subsets of tasks. These duty area aggregate variables could then be modeled.

In summary, multivariate analysis techniques and duty area aggregation provide other alternatives for reducing the dimensionality of the response data. The attractiveness of such alternatives is that they may not be subject to the same degree of information loss as in the case of averaging all the task-level data for an individual task into a single measure.

5.2.4 Developing Prediction Models. Recall that the regression models developed in this thesis accounted for at most 16% of the variance in the response, PC (maximum $R^2 = .16$). This means that at least 84% of the variance in the response remains unexplained by the developed models. To put this in context, consider Figure 27. Figure 27 indicates that there is a relatively enormous portion of variance in the response which remains to be explained. Recall that these results were comparable for previous PC studies (5) (13) (38). This means that there is probably significant improvement to be made in all phases of the job performance model development process.

A likely place to start improving the development of such models is in the job performance measurement realm. But, as has been proven over time, it is extremely difficult to develop a sound yet cost efficient system for collecting valid and reliable job performance data. This problem has been so pervasive in Industrial/Organizational Psychology that it has earned the fear-instilling name *the criterion problem*. Volumes have been written on job performance measurement and the

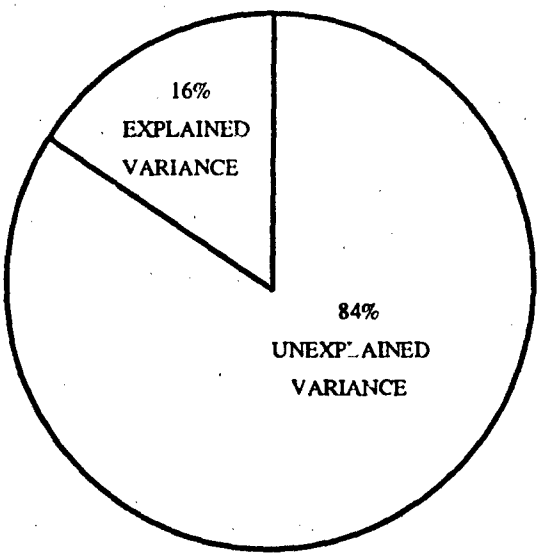


Figure 27. Pie Chart Representing the Explained vs. Unexplained Variance in Productive Capacity Given the Current Models

criterion problem. The topic of job performance measurement cannot be given the attention it is due in this limited space, thus the reader is referred to Reference (6) for an introduction to the topic.

Another likely area to be considered when seeking to improve job performance models is the predictor arena. Only two potential predictors were considered in the current effort, aptitude and experience. As with job performance measurement, volumes have been written concerning the relationship between numerous predictor variables and job performance. But, remember that PC is a fairly unique job performance measure in that it is supposed to measure a worker's capacity to produce, not how much he or she actually produces. This implies that many of the personality traits which would be expected to influence productivity would not be expected to influence PC. Such measures include worker motivation, job interest, work environment, and job satisfaction.

There still remain numerous potential predictors which would be expected to influence PC. These include the type and amount of technical school training, the type and amount of on-the-job training (OJT), the availability and quality of written technical guidance, the amount of

technical interaction with highly-skilled individuals, trouble-shooting and diagnostic ability, and general mental ability, just to name a few. Many such predictors could be considered for inclusion into Air Force job performance models. They may perhaps help to explain additional variance in the response.

A final area for model improvement might be the type of model itself. Perhaps linear regression-based models are simply insufficient for modeling the job performance of human beings. Humans are obviously highly complex entities with each being motivated and affected by countless factors. Added to this, the countless factors each influence different people in different ways. For these reasons alone, linear regression models may never be able to explain the majority of the variance in job performance.

In summary, there is significant improvement to be made in job performance modeling. Possible improvements could be made by improving the validity and reliability of the response (job performance measures), by considering other potential predictors and by considering different types of mathematical or maybe even non-mathematical models.

Appendix A. 454X1 Tasks Studied Under the Productive Capacity Project

Table 22: 454X1 Tasks Studied Under the Productive Capacity Project

Task	Description
E120	Make entries on supply issue and turn-in forms.
E143	Make entries on AFTO Form 350 (Reparable Item Processing Tag).
F153	Perform aircraft support air-conditioner visual and service inspection.
F154	Perform an aircraft support generator service inspection.
F155	Perform a service inspection on a load bank.
F157	Perform bomblift visual and service inspection.
F162	Perform a service inspection on a hydraulic test stand.
G171	Perform aircraft support air compressor periodic inspection.
G179	Perform combustor cap portion of a gas turbine compressor periodic inspection.
G181	Perform hydraulic test stand periodic inspection.
H202	Fabricate wiring.
H203	Isolate malfunction within electrical circuitry other than integrated or solid state.
H209	Measure resistance in AGE electrical systems by checking various circuits in the ignition system of the MC-2A.
H215	Perform AGE electrical systems operational checks.
H236	Research T.O.s, charts, or diagrams for electrical maintenance instructions.
H237	Solder electrical system wiring.
H238	Cut an electrical system wire in half and splice it together into a circle, using one crimp-type splice and one soldered heat shrink splice.
I247	Adjust distributor points.
I248	Adjust reciprocating engine fuel system components.
I251	Adjust turbine engine fuel system components.
I255	Change the generator in an NF-2.
I260	Clean commutator and slip rings on the generator of the NF-2.
I264	Troubleshoot the NF-2 generator for the following symptoms of malfunctions: (1) the engine will not start when cranked, and (2) the engine starts but backfires at the carburetor.
I275	Remove or install a carburetor on an MC-2A gasoline engine.
I283	Remove and install engine exhaust manifold, seals, gaskets, and common hardware.
I284	Remove and replace an alternator belt.
I286	Remove and install engine fuel pumps on the NF-2.
I299	Remove and install engine.
I300	Replace the flare fitting on a fuel line.
J332	Isolate the possible heater system malfunctions associated with a discrepancy that reads "burner will not ignite."
J340	Remove the burner control valve from an AGE heater.
J347	Remove and install heater engine.
J355	Remove and install temperature selector valve.
L406	Isolate hydraulic systems malfunction.

Table 22: (continued)

Task	Description
L421	Remove and install hydraulic lines on B-1 stand.
L436	Replace O-rings in hydraulic systems component.
L437	Research T.O.s, charts, or diagrams for AGE hydraulic systems maintenance.
M444	Assemble bleed air hose.
M446	Troubleshoot the MC-1A compressor for the discrepancy "Compressor fails to unload at 3600 psi."
M447	Perform AGE pneumatic system operational check.
N475	Isolate brake sy .tem malfunction.
N477	Repack wheel bearings of one wheel on AGE equipment (NF-2).
N486	Remove and install AGE brake pads.
N487	Remove and install AGE fuel tank.
N488	Change an AGE tire and tube assembly.
N494	Remove and install one six inch bolted hinge.
N503	Look up the part number, source code, and work unit code to requisition a new axle assembly for an MC-2A compressor (with date plate containing the following information: MFG- Davey Compressor Company, Contract #-DSA 700-74C-9004, Serial #-16160, Reg #-4310-75-D18-6160, Model #-2MC-2, Part #-27391).
P549	Perform an operator's inspection of an AF vehicle, completing AFTO Form 373.
P554	Pick up and deliver -60.
P555	Prepare AGE (NF-2) for shipment during a training exercise or mobilizatien.

The above task descriptions were taken from Reference (24)

Bibliography

1. Alley, William E. and Teachout, Mark S. "Aptitude and Experience Trade-Offs on Job Performance," *Proceedings of the 98th Annual Convention of the American Psychological Association.* 1990.

2. Avinger, Charles R. *Analysis of Learning Curve Fitting Techniques.* MS Thesis, AFIT/GSM/LSQ/87S-3. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB, OH, September 1987.

3. Bentley, Barbara A., Ringenbach, Kathleen L. and Augustin, James W. *Development of Army Job Knowledge Tests for Three Air Force Specialties, Interim Technical Paper AFHRL-TP-88-11, February 1987-December 1987.* Contract F41689-86-D-0052. Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory, May 1989 .

4. Box, George E. P. and Draper, Norman R. *Empirical Model-Building and Response Surfaces.* New York: John Wiley & Sons, Inc., 1987.

5. Carpenter, Michael A., Monaco, Salvatore J., O'Mara, Francis E. and Teachout, Mark S. *Time to Job Proficiency: A Preliminary Investigation of the Effects of Aptitude and Experience on Productive Capacity, Final Technical Paper AFHRL-TP-88-17, January 1986-July 1988.* Contract F41689-84-D-0002. Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory, June 1989 .

6. Cascio, Wayne F. *Applied Psychology in Personnel Management* (Third Edition). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1987.

7. Demetriades, Ernest T., Captain, USAF. "Productive Capacity." Address to Wayne S. Sellman, Director for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel), The Pentagon, Washington DC, April 1992.

8. Department of the Air Force. *Enlisted Personnel.* AFR 39-1. Washington DC: HQ USAF, 15 March 1991.

9. Department of the Air Force. *Vocational Interests for Career Enhancement (VOICE) Form D.* AFPT 600. Washington DC: HQ USAF, 1 June 1987.

10. Department of Defense. *Armed Services Vocational Aptitude Battery (ASVAB) Test Manual for Forms 8, 9, 10, 11, 12, 13, and 14.* DoD 1304.12AA. North Chicago, IL: United States Military Entrance Processing Command, 1 July 1984.

11. Dillon, William R. and Goldstein, Matthew. *Multivariate Analysis Methods and Applications.* New York: John Wiley & Sons, Inc., 1984.

12. Fairbanks, Benjamin, A., Jr. "A Revalidation of the U.S. Air Force Vocational Interest Career Examination." Unpublished Technical Report. Air Force Human Resources Laboratory, Brooks AFB, TX, July 1983.

13. Faneuff, Robert S., Valentine, Lonnie D., Jr., Stone, Brice M., Curry, Guy L. and Hageman, Dwight C. *Extending the Time to Proficiency Model for Simultaneous Application to Multiple Jobs, Interim Technical Paper AFHRL-TP-90-42, October 1988-April 1990.* Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, July 1990.

14. Fox, William M. "The Improved Nominal Group Technique (INGT)," *Journal of Management,* 8: 20-27 (1989).

15. Hedge, Jerry W. and Lipscomb, M. Suzanne. *Walk-Through Performance Testing: An Innovative Approach to Work Sample Testing. Interim Technical Paper AFHRL-TP-87-8, May 1983-August 1984.* Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory, September 1987.

16. Hedge, Jerry W. and Teachout, Mark S. *Job Performance Measurement: A Systematic Program of Research and Development, Interim Technical Paper AFHRL-TP-86-37, June 1985-April 1986.* Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory, November 1986 .

17. Hosmer, David W., Jr. and Lemeshow, Stanley. *Applied Logistic Regression.* New York: John Wiley & Sons, 1989.

18. Jordan, Raymond B. *How to Use the Learning Curve.* Boston: Farnsworth Publishing, Inc., 1965.

19. Laue, Francis J., Teachout, Mark S. and Harville, Donald L. *Assessing the Substitutability of Surrogate Measures of Job Performance for Hands-On Work Sample Tests, Final Technical Paper AL-TP-0025, January 1989-January 1990.* Contract F41869-86-D-0052. Brooks AFB, TX: Technical Training Research Division, Human Resources Directorate, Armstrong Laboratory, June 1992.

20. Law, Averill M. and Kelton, W. David. *Simulation Modeling and Analysis.* New York: McGraw-Hill Book Company, 1982.

21. Leighton, Daniel L., Kageff, Linda L., Mosher, Gregory P., Gribben, Monica A., Faneuff, Robert S., Demetriades, Ernest T. and Skinner, M. Jacobina. *Measurement of Productive Capacity: A Methodoloyy for Air Force Enlisted Specialties, Interim Technical Paper AL-TP-1992-0029, September 1990-September 1991.* Contract F49642-86-D-0001. Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory, June 1992.

22. Liebman, Judith, Lasdon, Leon, Schrage, Linus and Waren, Allan. *Modeling and Optimization with GINO.* Palo Alto: The Scientific Press, 1986.

23. Lipscomb, M. Suzanne and Hedge, Jerry W. *Job Performance Measurement of Air Force Enlisted Personnel, Interim Technical Paper AFHRL-TP-87-58, March 1987-July 1987.* Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory, June 1988.

24. Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory. "Aerospace Ground Equipment (AGE) Specialty, AFS 454X1, Tasks For Time Estimation." Task booklets used in the Air Force's Productive Capacity Project data collection, Brooks AFB, TX. March 1991

25. Mendenhall, William, Wackerly, Dennis D. and Schaeffer, Richard L. *Mathematical Statistics With Applications* (Fourth Edition). Boston: PWS-Kent Publishing Company, 1990.

26. Nachmias, David and Nachmias, Chava. *Research Methods in the Social Sciences* (Third Edition). New York: St. Martin's Press, 1987.

27. Neter, John, Wasserman, William and Kutner, Michael H. *Applied Linear Statistical Models* (Third Edition). Homewood, IL: Richard D. Irwin, Inc., 1990.

28. Neter, John, Wasserman, William and Kutner, Michael H. *Applied Linear Regression Models* (Second Edition). Homewood, IL: Richard D. Irwin, Inc., 1989.

29. Niebel, Benjamin W. *Motion and Time Study* (Fifth Edition). Homewood, IL: Richard D. Irwin, Inc., 1972.

30. Office of the Assistant Secretary of Defense (Force Management and Personnel). *Joint-Service Efforts to Link Military Enlistment Standards to Job Performance.* Report to the House Committee on Appropriations. April 1992.

31. Ree, Malcolm J. and Earles, James A. *Subtest and Composite Validity of ASVAB Forms 11, 12, and 13 for Technical Training Courses, Interim Technical Report AL-TR-1991-0107, January*

1990–August 1991. Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory, February 1992.

32. Ringenbach, Kathleen L. "Development of a Generalized Motivation Scale," *Proceedings of the Military Testing Association 31st Annual Conference*. 1989.

33. Ringenbach, Kathleen L. "Development of a Generalized Motivation Scale." Unpublished Technical Report. Air Force Human Resources Laboratory, Brooks AFB, TX, January 1989.

34. Sackett, Paul R., Zedeck, Sheldon, and Fogli, Larry. "Relations Between Measures of Typical and Maximum Job Performance," *Journal of Applied Psychology*, Vol.73, No.3: 482–486 (1988).

35. Schmidt, Frank L., Hunter, John E. and Outerbridge, Alice N. "Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance," *Journal of Applied Psychology*, Vol.71, No.3: 432–439 (1986).

36. Schmidt, Frank L., Hunter, John E., Outerbridge, Alice N. and Goff, Stephen. "Joint Relation of Experience and Ability With Job Performance: Test of Three Hypotheses," *Journal of Applied Psychology*, Vol.73, No.1: 46–57 (1988).

37. Siegel, Laurence and Lane, Irving M. *Psychology in Industrial Organizations* (Third Edition). Homewood, IL: Richard D. Irwin, Inc., 1974.

38. Skinner, M. Jacobina. Regression analysis results and summary statistics computed using the SAS data analysis software. Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory, January 1993.

39. Skinner, M. Jacobina, Faneuff, Robert S. and Demetriades, Ernest T. "Developing Benchmarks to Scale Task Performance Times," *Proceedings of the Military Testing Association 33rd Annual Conference*. 1991.

40. United States Air Force. *Occupational Survey Report, Aerospace Ground Equipment, AFSC 454X1*. AFPT 90-454-904. Randolph AFB, TX: Occupational Measurement Squadron, Air Training Command, January 1992.

41. Wright, T. P. "Factors Affecting the Cost of Airplanes," *Journal of the Aeronautical Sciences*, Vol.3, No.4: 122–128 (February 1936).

Vita

Captain Robert Faneuff was born in Toledo, Ohio on 26 October 1964. He was raised in Rossford, Ohio, a Toledo suburb. He graduated from Rossford High School in 1983 and attended the United States Air Force Academy immediately after high school. He graduated from the Air Force Academy in 1987 with a Bachelor of Science degree in Behavioral Science (specialty: Human Factors Engineering). His Air Force career includes a tour with the Human Resources Directorate, Armstrong Laboratory, Brooks Air Force Base, Texas. There he served as a Behavioral Scientist researching the service selection and job classification of Air Force enlisted personnel. He was responsible for designing, implementing, and analyzing experiments aimed at maintaining or improving the Air Force's enlisted force acquisition system. He authored or coauthored six technical papers and conference papers documenting research he had accomplished. Following his assignment at Brooks Air Force Base, he was assigned to the Air Force Institute of Technology (AFIT), School of Engineering, to pursue a Master of Science degree in Operations Research. Following graduation from AFIT, he will be assigned to the Air Force Quality Center, Maxwell Air Force Base, Alabama.

Permanent address: 5276 Access Rd.
Dayton, Ohio 45431

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1993	Master's Thesis

4. TITLE AND SUBTITLE
PREDICTING THE PRODUCTIVE CAPACITY OF AIR FORCE AEROSPACE GROUND EQUIPMENT PERSONNEL USING APTITUDE AND EXPERIENCE MEASURES

5. FUNDING NUMBERS

6. AUTHOR(S)
Robert S. Faneuff, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GOR/ENS/93M-05

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AL/HRM
Brooks AFB
San Antonio, TX 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This study investigated the effects of mechanical aptitude and job experience on the job performance of 204 Air Force Aerospace Group Equipment (AGE) mechanics. Job performance was expressed as *productive capacity (PC)*, which is derived from estimated performance times on job tasks. PC measures were derived for 50 tasks typically performed by airmen in the specialty. Aptitude measures took the form of Mechanical percentile composite scores on the Armed Services Vocational Aptitude Battery (ASVAB). A second-order logistic model was used to regress PC on aptitude and experience at the task level and at the overall job, or aggregate, level. Model R^2s were generally low. For the tasks, R^2s ranged from .01 to .13, and for the aggregate model the R^2 was about .16. Generally, experience was a significant predictor but aptitude was not. There was also no indication of an aptitude/experience interaction. These results were verified through forward stepwise regression. There was some evidence that airmen may experience some skill degradation on production-type tasks at around the six year point as they transition to supervisory roles.

14. SUBJECT TERMS
Job Performance, Productive Capacity, Logistic Regression, ASVAB

15. NUMBER OF PAGES
139

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRAC
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL